DISCOVERY OF EARLY CANCER SIGNATURES IN CIRCULATING RNA AND EXTRACELLULAR VESICLES

By

Hyun Ji Kim

A DISSERTATION

Presented to the Department of Biomedical Engineering
of the Oregon Health & Science University
School of Medicine

in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
in Biomedical Engineering

June 2021

Department of Biomedical Engineering
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph.D. Dissertation of

Hyun Ji Kim

"Discovery of Early Cancer Signatures in Circulating RNA and Extracellular Vesicles"

Has been approved

Mentor: Thuy Ngo, Ph.D.
Professor of Biomedical Engineering

Member/Chair: Terry Morgan, M.D., Ph.D.
Professor of Pathology and Biomedical Engineering

Member: Sadik Esener, Ph.D.
Professor of Biomedical Engineering

Member: Reid Thompson, M.D., Ph.D.
Professor of Biomedical Engineering

Member: Stephen Quake, Ph.D.
Professor of Bioengineering, Stanford University

Dedication

I would like to dedicate this work to my family and people struggling with cancer, cancer survivors, and their families

Acknowledgements

First of all, I would like to thank all the commitment of the Knight Cancer Challenge and fund raising efforts in pioneering work for early cancer detection. In 2013, Nike co-founder Phil Knight and his wife, Penney, pledged to donate $500 million to the Knight Cancer Institute if OHSU could raise an additional $500 million. To do this, Dr. Sadik Esener, my mentor from University of California, San Diego, who is now the director of the Cancer Early Detection Advanced Research center, aimed at saving lives by developing advanced technologies to detect cancer early. In pioneering translational research at Oregon Health and Science University, I met my mentor, Dr. Thuy Ngo, who I respect for all her guidance and appreciate all the powers and energies to do science. I give special thanks to my mentor for giving me the great chance to be a part of her group, believing in me, and teaching me to be a scientist. She taught me how to develop next generation sequencing tests by discovering new biomarkers through extracting cell-free RNA from human plasma. Her great mentorship helped me to achieve my full potentials. Besides my mentor, I would like to thank the rest of my dissertation committee, Dr. Terry Morgan, Dr. Sadik Esener, and Dr. Reid Thompson for their great support and invaluable advice. I am thankful to Dr. Terry Morgan for serving as my committee chair and guiding me to complete my PhD. I am honored that Dr. Stephen Quake from Stanford could serve as the final member of my committee.

I also would like to thank all the teams that have been helping me in various aspects through multiple projects. Dr. Owen McCarty, Dr. Samuel Tassi Yunga, and Dr. Anh Ngo provided platelet expertise to understand the blood processing impact on cell free

messenger RNA and extracellular vesicles. I am grateful to Breeshey Roskams-Hieter who taught me Linux-based pipelines and R to run RNA-seq analysis and Parvana Anur who helped building a data pipeline for the cell free RNA project. Elias Spiliotopoulos and Ward Kirschbaum helped to complete cell free RNA validation with qPCR, and Rowan Callahan helped with tissue deconvolution. Dr. Josiah Wagner helped me to investigate the diurnal stability of DNA, RNA, and extracellular vesicles. I also would like to thank Dr. Fehmi Civitci for his helpful discussion and critical thinking. Matthew Rames and Dr. Josephine Briand helped me with immunoprecipitation. I am especially thankful to Matthew Rames who always encouraged me to work hard, brainstorm ideas together, and helped me to extract 168 high-quality RNA from size-fractionated plasma. I also had great project advisors: Gene Tu, Dr. Bruce Branchaud, and Dr. Paul Spellman. Dr. Jeong Yoon Lim provided a great biostatistics support. I also had the pleasure to work alongside other members in CEDAR, and thank CEDAR project management teams and milestone committees for their guidance.

The thesis would not have come to a successful completion without the help I received from all my collaborators. I also thank Dr. Timothy Butler, Christopher Thomas Boniface, and Dr. Nicholas Wang who taught me how to use chip-based capillary electrophoresis to analyze RNA and 386 well qPCR. Dr. Terry Morgan, Mayu Morita, Pamela Canaday, Brianna Garcia, Dorian Latocha, and Randall Armstrong for their flow cytometry expertise. During the course of flow cytometry standardization, I was fortunate to cross paths with Dr. Joshua A. Welsh from NIH whose discussion and support were truly inspiring and motivated my project. I also thank Dr. Claudia Lopez, director of the Multiscale Microscopy Core, and Matthew Rames for Transmission Electron Microscopy

experience. I also thank all the clinical collaborators; Dr. Reid Thompson (Lung Cancer), Dr. Dove Keith and Dr. Rosalie Sears (Pancreatic Cancer), Dr. Liana Tsikitis (Colorectal cancer), Dr. Scott Naugler (Liver Cancer and Cirrhosis), Dr. Emma Scott and Dr. Gullu Gorgun (Multiple Myeloma and Monoclonal Gammopathy of Undetermined Significance), and Zach Stupor and Dr. Chris Corless from all cancer types. I thank biorespository teams; Katie C. Johnson-Camacho, Dorien Hartunian and Meghan Fitzgerlad. I truly appreciate all the patients and healthy volunteers who donated their blood to science.

Lastly, but certainly not least, I would like to thank my family for their endless love, long waiting, and great support through this journey.

Table of Contents

Dedication ..... **iii**
Acknowledgements ..... **iv**
Table of Contents ..... **vii**
List of Tables ..... **x**
List of Figures ..... **xii**
List of Abbreviations ..... **xvi**
Abstract ..... **xviii**
Chapter I: Introduction to Biomarkers for Early Cancer Detection **1**
1.1 Introduction 2
1.2 Fundamentals of extracellular vesicles 3
1.3 EV isolation, quantification, and characterization techniques 7
1.3.1 EV isolation 7
1.3.1.1 Ultracentrifugation 7
1.3.1.2 Filtration and size-exclusion chromatography 8
1.3.1.3 Polymer precipitation and affinity based bead capture 9
1.3.1.4 Asymmetric-flow field fractionation 9
1.3.2 EV quantification and characterization 12
1.3.2.1 Biophysical property measurement of EVs 12
1.3.2.2 Protein Characterization of EVs 15
1.3.2.3 RNA characterization of EVs 21
1.4 Fundamentals of cell-free RNA 23
1.5 RNA isolation, quantification and characterization techniques 25
1.5.1 RNA isolation 26
1.5.1.1 RNA extraction 26
1.5.2 RNA quantification 30
1.5.2.1 RNA quantification by UV absorbance 30
1.5.2.2 Fluorescence measurement using nucleic acid dye 31
1.5.2.3 qPCR mechanism and detection 32

1.5.3 RNA sequencing, alignment, and characterization 36
1.5.3.1 RNA sequencing 36
1.5.3.2 Workflow for RNA sequencing 38
1.5.3.3 Normalization 39
1.5.3.4 Downstream analysis 42
1.6 Gaps in current understanding & layout of the thesis 42
Chapter II: Irreversible alteration of extracellular vesicle and cell-free messenger RNA profiles in human plasma associated with blood processing and storage 44
2.1 Abstract 45
2.2 Introduction 46
2.3 Materials and Methods 48
2.4 Results 52
2.5 Discussion 63
2.6 Conclusions 66
Chapter III: Diurnal stability of cell-free DNA, cell-free RNA, and extracellular vesicles in human plasma samples 69
3.1 Abstract 70
3.2 Introduction 70
3.3 Materials and Methods 73
3.4 Results and Discussion 79
3.5 Conclusions 88
Chapter IV: Plasma cell-free RNA profiling enables multiclass pan-cancer detection and distinguishes cancer from pre-malignant conditions 91
4.1 Abstract 92
4.2 Introduction 92
4.3 Materials and Methods 95
4.4 Results 99
4.5 Discussion 109
4.6 Conclusions 111
Chapter V: Selective packaging of extracellular vesicles RNA association with cancer progression 114

5.1 Abstract 115
5.2 Introduction 115
5.3 Materials and Methods 119
5.4 Results 124
5.5 Discussion 137
5.6 Conclusions 140
Appendix A: Supplementary Tables and Figures 142
References 180

List of Tables

Table 1.1 | Summary of different EV isolation methods. 11

Table 3.1 | Age and sex of the four healthy donors (HDs) volunteered for this study. 73

Table 3.2 | Summary of the permutation tests comparing the two draw days, the five draws across the day, or individuals to determine significant sources of variation.. 82

Supplementary Table S3.1 | Total plasma cfDNA concentration summaries and statistics as measured by Qubit. 152

Supplementary Table S3.2 | Plasma *TERT* concentration summaries and statistics as measured by ddPCR. 153

Supplementary Table S3.3 | Plasma *NAGK* concentration summaries and statistics as measured by ddPCR. 154

Supplementary Table S3.4 | Total plasma cfRNA concentration summaries and statistics as measured by Bioanalyzer. 155

Supplementary Table S3.5 | Plasma *ACTB* cDNA concentration summaries and statistics as measured by ddPCR. 156

Supplementary Table S3.6 | Plasma *GAPDH* cDNA concentration summaries and statistics as measured by ddPCR. 157

Supplementary Table S3.7 | Plasma CD81+ EV count summaries and statistics as measured by flow cytometry. 158

Supplementary Table S3.8 | Plasma CD63+ EV count summaries and statistics as measured by flow cytometry. 159

Supplementary Table S3.9 | Plasma CD41+ EV count summaries and statistics as measured by flow cytometry. 160

Supplementary Table S3.10 | Plasma CD9+ EV count summaries and statistics as measured by flow cytometry. 161

Supplementary Table S4.1 | Summary of input reads, unique reads, exon fraction, intron fraction, intergenic fraction, and protein coding fraction. 162

List of Figures

Figure 1.1 | Schematic of EV biogenesis and contents of EVs. 4

Figure 1.2 | Function and roles of EVs. 6

Figure 1.3 | Asymmetric field flow size fractionation of EVs. 10

Figure 1.4 | TEM image of transferrin receptor containing extracellular vesicles from reticulocytes. 13

Figure 1.5 | qNano instrument and mode of operation 14

Figure 1.6 | Analysis of platelet-derived EVs using flow cytometer 17

Figure 1.7 | Proteomic analysis of EVs from surgically removed pancreatic cancer tissue explant 20

Figure 1.8 | ExRNA atlas from Extracellular RNA Communication Consortium 22

Figure 1.9 | Characterization of miRNA stability and detection of human prostate cancer by serum levels 24

Figure 1.10 | Total RNA isolated using silica carbide compared to other RNA extraction kits 28

Figure 1.11 | Hierarchical clustering analysis of miRNA among different biofluid types and exRNA isolation methods 29

Figure 1.12 | Overview of qPCR workflow for measuring RNA using SYBR green. 33

Figure 1.13 | Overview of RT-qPCR workflow for measuring RNA using TaqMan. 34

Figure 1.14 | Basic principle of PCR and relative fluorescence cycle number 35

Figure 1.15 | Schematic of SMARTer stranded total RNAseq library preparation 37

Figure 1.16 | Workflow of RNAseq analysis 39

Figure 1.17 | RNAseq normalization and interpretation of expression data 41

Figure 2.1 | Light scattering calibration 54

Figure 2.2 | Fluorescence calibration 55

Figure 2.3 | Effect of differential centrifugation on EVs using flow cytometry 57

Figure 2.4 | Effect of freeze thaw cycle on EVs using flow cytometry 59

Figure 2.5 | Effect of post-thaw processing on EVs using flow cytometry 60

Figure 2.6 | Effect of freeze thaw and post-thaw processing on cf-mRNAs using qRT-PCR 62

Figure 3.1 | Schematic of the HD sampling procedure used to obtain plasma for analysis. 74

Figure 3.2 | Abundance of plasma-derived cfDNA across the five sampled time points. 81

Figure 3.3 | Abundance of plasma-derived cfRNA across the five sampled time points. 84

Figure 3.4 | Abundance of plasma-derived EVs across the five sampled time points. 87

Figure 4.1 | cfRNA profiles distinguish between cancer vs. healthy donors. 100

Figure 4.2 | cfRNA profiles enable classification of pan-cancers. 104

Figure 4.3 | cfRNA profiles distinguish between healthy, MGUS and multiple myeloma donors: 106

Figure 4.4 | cfRNA profiles distinguish between healthy, liver cirrhosis and liver cancer donors: 108

Figure 5.1 | Characterization of distinct EV subtypes through plasma fractionation 125

Figure 5.2 | Transcriptomic Analysis of EVs and Non-vesicles 128

Figure 5.3 | Relative quantification of RNA by qRT-PCR and immunoprecipitation 130

Figure 5.4 | Distinct cancer differentiating cell-free mRNA across fractions in human plasma 132

Figure 5.5 | Specific cfRNA signatures associated with high risk group and cancer 135

Supplementary Figure S2.1 | Flow cytometry experimental assay controls 143

Supplementary Figure S2.2 | Freeze thaw effect and plasma EV detergent treatment ... 144

Supplementary Figure S3.1 | Total number of droplets accepted by the QX200 ddPCR droplet reader for nucleic acid quantitation. 145

Supplementary Figure S3.2 | Negative controls for ddPCR measurement of cfDNA. ... 146

Supplementary Figure S3.3 | Nonparametric spearman correlations ... 147

Supplementary Figure S3.4 | Total number of droplets accepted by the QX200 ddPCR droplet reader for nucleic acid quantitation. ... 148

Supplementary Figure S3.5 | Negative controls for ddPCR measurement of cDNA derived from cfRNA. ... 149

Supplementary Figure S3.6 | Flow cytometry set-up and gating for detection of plasma EVs. ... 150

Supplementary Figure S3.7 | Flow cytometry assay controls for plasma EV measurement. ... 151

Supplementary Figure S4.1 | Distribution of sequencing reads across all 71 samples. .. 163

Supplementary Figure S4.2 | Distribution of exon/intro and intergenic fractions across all 71 samples. ... 164

Supplementary Figure S4.3 | Coverage of the transcriptome across all 71 samples ... 165

Supplementary Figure S4.4 | Volcano plots from cfRNA pairwise cohorts ... 166

Supplementary Figure S4.5 | Differential gene expression permutation tests ... 167

Supplementary Figure S4.6 | cfRNA Gene Ontology ... 168

Supplementary Figure S5.1 | Description of input reads, unique reads, exon, intron and intergenic fraction ... 169

Supplementary Figure S5.2 | Distribution of exon, intron, and intergenic fractions across fractions ... 170

Supplementary Figure S5.3 | Summary of proportion of transcript types across all samples ... 171

Supplementary Figure S5.4 | Transcriptomic analysis of EVs and non-vesicles per condition ... 172

Supplementary Figure S5.5 | Lung cancer associated genes packaged in protein enriched fraction ... 173

Supplementary Figure S5.6 | Liver cancer associated genes across fractions in human plasma 174

Supplementary Figure S5.7 | Multiple myeloma associated genes across fractions in human plasma 175

Supplementary Figure S5.8 | Intersection of cancer distinguishing genes across plasma fractions..... 176

Supplementary Figure S5.9 | Gene set enrichment for cancer distinguishing genes 177

Supplementary Figure S5.10 | Gene set enrichment analysis associated with healthy, liver cirrhosis, and liver cancer comparisons 178

Supplementary Figure S5.11 | Gene ontology analysis associated with healthy, MGUS, and multiple myeloma comparisons 179

List of Abbreviations

AF4	Asymmetric flow field flow fractionation
CCD	Charge coupled device
cDNA	Complementary DNA
cf-RNA	Cell free RNA
ddPCR	Digital droplet PCR
DE	Differentially expressed
dsDNA	Double stranded DNA
ERCC	Extracellular RNA communication consortium
EV	Extracellular vesicle
EV-RNA	Extracellular vesicle RNA
FSC	Forward scatter
GO	Gene ontology
HCC	Hepatocellular carcinoma
HPLC	High performance liquid chromatography
HR-MS	High resolution mass spectroscopy
IA	Immunoaffinity
LC	Liver cirrhosis
LDA	Linear discriminant analysis
LOOCV	Leave one out cross validation
LVQ	Learning vector quantization
MESF	Molecules of equivalent soluble fluorochrome
MF	Membrane filter
MGUS	Monoclonal gammopathy of undetermined significance
miRNA	Micro RNA
MM	Multiple myeloma
mRNA	Messenger RNA
NGS	Next generation sequencing
NIST	National institutes of standards and technology

NTA	Nanoparticle tracking analysis
PCA	Principal component analysis
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
RF	Random forest
RLE	Relative log expression
ROC	Receiver operating characteristic
RPKM	Reads per kilobase per million mapped reads
RPM	Reads per million mapped reads
RT	Reverse transcriptase
RT-qPCR	Real time quantitative polymerase chain reaction
SEC	Size exclusion chromatography
SiC	Silica carbide
SiF	Silica fiber
SiM	Silica membrane
SSC	Side scatter
ssDNA	Single stranded DNA

Abstract

This thesis work highlights development and applications for both circulating extracellular vesicles (EV) and cell-free RNA (cf-RNA) toward liquid biopsy based early cancer detection methodologies. Effective detection and monitoring for signatures of oncological disease in a noninvasive manner are urgently needed to reduce the morbidity and mortality caused by cancer. Circulating EVs and cf-RNA are intensely sought after biomarkers in liquid biopsy. Their roles in cell-to-cell communication, ability to reflect phenotypic changes from cells, and tissues of origin are becoming better understood. Despite this, current literatures present key challenges which limit the promise of liquid biopsy based early cancer detection: i) there is a lack of standardized blood processing for multi-omics which minimizes ex-vivo processing artefacts via discerning true in-vivo signatures from ex-vivo artefacts; ii) daily fluctuations and the influence of meal consumption on EV and cf-RNA levels are not clear; iii) a comprehensive study of cf-RNA for cancer detection, pan cancer discernment, and high risk group identification has not been conducted; and iv) the selective packaging of cf-RNA carriers and its association with cancer are unknown.

The chapter one is an overview of existing technologies commonly utilized in EV and cf-RNA studies and includes considerations for complex biofluids. In chapter two, we systematically evaluated the effect of preanalytical variation on the yield and purity of EVs and cf-RNA in human plasma. Notably, we found that centrifugation and temperature resulted in the highest EV and cf-RNA variability, owing to the release of ex-vivo derived EVs from platelets. The extent of these technical artefacts significantly differed for distinct

EV sizes and types of cf-RNA transcripts, highlighting the importance to minimize sources of technical artefacts. In chapter three, we assessed the diurnal and interpersonnel variation on EVs and cf-RNA, which may impact biomarker discovery. Through serially sampling a preliminary cohort, we showed that EV and cf-RNA were consistent over time for a given individual. In contrast, we found a significant interpersonal variation, highlighting the importance in larger population screening and understanding of person-to-person variation.

In chapter four and five, we investigated the potential clinical utility and biological roles of EVs and cf-RNA as noninvasive biomarkers for early cancer detection. In chapter four, we revealed cell-free messenger RNA (cf-mRNA) transcripts not only differentiate the presence of cancer, but also classified individual cancer types and high-risk groups using cf-RNA sequencing and machine-learning approaches. However, how these signatures are protected from the RNase-rich environment in plasma and how cancer may dysregulate cf-mRNA signatures remained unknown. Therefore, in chapter five, we aimed to determine if EVs are the major cf-mRNA carrier and how to identify which cf-mRNA transcripts are dysregulated in different cancer types and high-risk groups. To address the role of cf-RNA packaging as a novel biomarker, we sequenced the RNA of EVs and non-vesicles from size fractionated human plasma. Critically, we found the majority of cf-mRNA were contained within EV-enriched plasma fractions, while also being protected from RNase digestion. In addition, we discovered distinct cancer and high-risk group distinguishing genes were selectively packaged across plasma fractions. Ultimately, these specific gene sets reflected an imbalance of secreted RNA found in cancer progression as a form of cell-to-cell communication.

Chapters two, three, four, and five are, at least in part, reprints of the following publications:

1. **Hyun Ji Kim**, Matthew Rames, Samuel Tassi Yunga, Randall Armstrong, Mayu Morita, Owen McCarty, Fehmi Civitci, Terry Morgan, and Thuy T. M. Ngo, "Irreversible alteration of extracellular vesicle and cell-free messenger RNA profiles in human plasma associated with blood processing and storage", Submitted to *Scientific Reports* (2021).

2. Josiah T. Wagner, **Hyun Ji Kim**, Katie C. Johnson-Camacho, Taylor Kelley, Laura F. Newell, Paul T. Spellman, and Thuy T. M. Ngo, "Diurnal stability of cell-free DNA and cell-free RNA in human plasma samples" *Scientific Reports* volume 10, Article number: 16456 (2020).

3. **Hyun Ji Kim**, Breeshey Roskams-Hieter, Pavana Anur, Josiah T. Wagner, Fehmi Civitci, Paul Spellman, Reid F. Thompson, Willscott E. Naugler, and Thuy T. M. Ngo, "Plasma cell-free RNA profiling enables multiclass pan-cancer detection and distinguishes cancer from pre-malignant conditions", In preparation for resubmission (2021).

4. **Hyun Ji Kim**, Breeshey Roskams-Hieter, Matthew Rames, Josephine Briand, Josiah Wagner, Aaron Doe, and Thuy T. M. Ngo, "Selective packaging of extracellular vesicles RNA association with cancer progression", In preparation (2021).

Chapter I: Introduction to Biomarkers for Early Cancer Detection

1.1 Introduction

Effective detection and monitoring for signatures of malignant disease in a noninvasive manner are urgently needed to reduce the morbidity and mortality caused by cancer. Early stage cancers are often asymptomatic, making early diagnosis difficult. However, routine blood tests which detect abnormal molecular signatures can help diagnose cancers earlier. Accordingly, identifying robust circulating signatures for pan-cancer detection is the “holy grail” for liquid biopsy, with an ever-expanding body of literature addressing it. Among the wide variety of circulating biomarker carriers being studied, extracellular vesicles (EVs) and circulating nucleic acids especially cell-free messenger RNA (cf-mRNA) are particularly sought after, as they have been shown to carry specific information from their cells of origin while also potentially reflect phenotypic changes.

Owing to significant progress made in identifying extracellular vesicles and cell-free nucleic acids signatures, improved isolation and analytical methods have aided in understanding the biology of extracellular vesicles spurring research into diagnostic applications for human biofluids. Despite significant progress made in understanding circulating EVs and cf-RNA, integration of existing methods to probe the utility of these biomarkers in human plasma has presented a number of challenges. Specifically, a lack of standardized protocols coupled with insufficiently sensitive detection and analytical tools has led to this gap in translation from model systems to more accurately quantifying these biomarkers from patient plasma.

My thesis work progresses the key steps needed to develop such early-cancer differentiating liquid biopsy biomarkers utilizing circulating EVs and cf-mRNA, as well as

a demonstration of how these signatures can be stratified for cancer detection. Transferring to OHSU from UC San Diego, where I began my PhD research into nanoparticle drug delivery, the focus of my work became how I could leverage my growing breadth of analytical experience to help push the boundaries of liquid biopsy based early cancer detection at the growing OHSU Knight Cancer Research Center. In this chapter, I describe the fundamentals of EVs and cf-RNA, followed by an overview of current methods and findings related to the isolation, identification, and quantification of these liquid biopsy biomarkers. To set the foundation for discussions in subsequent chapters, we will touch upon some of the current gaps in understanding in how to employ EVs and cf-RNA as robust liquid biopsy biomarkers. Ultimately, this provides the basis for the underlying hypothesis of my thesis.

1.2 Fundamentals of extracellular vesicles

Extracellular vesicles (EVs) are membrane-enclosed vesicles released by many cell types and found in every bodily fluid. In 1983, EVs were reported from sheep reticulocytes using transferrin receptor [1]. Iodine-125 or FITC labelled reticulocyte's transferrin receptors were externalized into the extracellular space [1]. Pan and Johnstone found multivesicular bodies were fused with plasma membrane, leading to the release of small vesicles under 100 nm in diameter [1]. EVs are typically defined by size, biogenesis, density (1.1.3 – 1.19 g/ml), and certain enriched protein markers. EVs are generated in a process that involves formation of multivesicular bodies and fusion with the plasma membrane (**Figure 1.1**). Size, heterogeneity and different biogenesis mechanisms divide EVs into microvesicles, exosomes, and exomeres. Microvesicles are typically 150 nm -

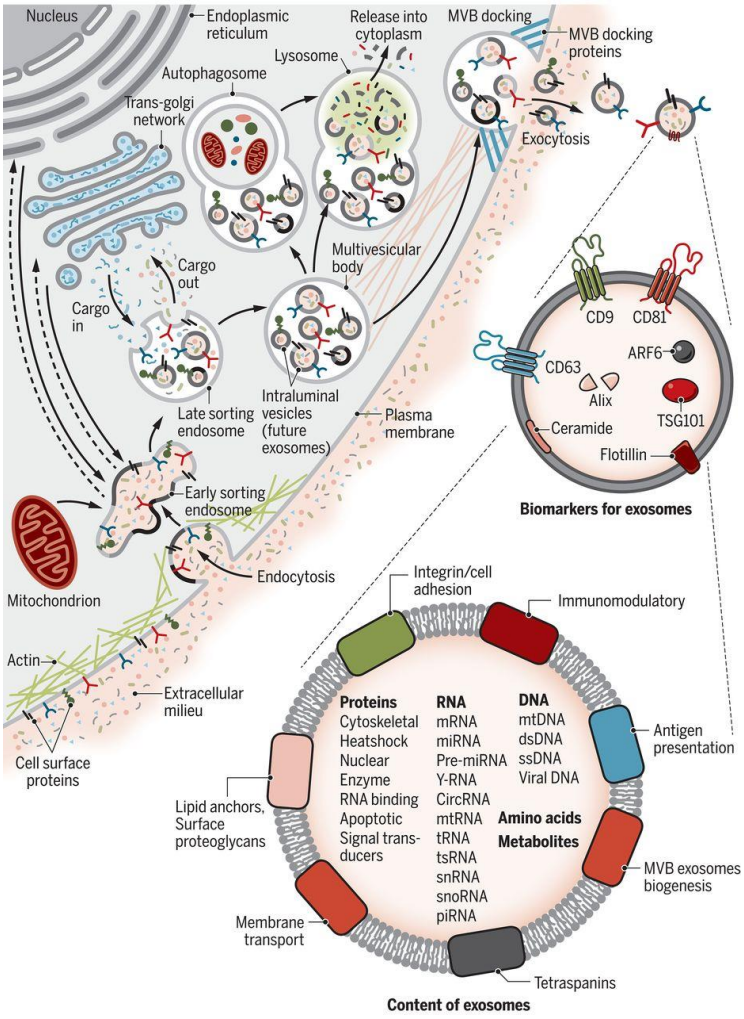


Figure 1.1 | Schematic of EV biogenesis and contents of EVs.

Through endocytosis extracellular contents enter the cells, where in plasma membrane invagination presents the outer membrane orientation towards outside. This budding process results in formation of early sorting endosome which gives rise to late sorting endosome. Second invagination in the late sorting endosome leads to generation of intraluminal vesicle, which proteins originally on the cell surface could distributed on the membrane. During which multivesicular body fuse with plasma membrane, exocytosis occurs and releases EVs to the outer membrane. Several proteins such as teraspanin markers (CD9, CD63, and CD81), TSG101, and Flotillin are common exosome markers. Intracellular proteins, RNA, DNA, and amino acids can also be found inside EVs. Reprinted with permission from [2]. Copyright 2020 American Association for the Advancement of Science.

1 μm in size and released via membrane blebbing from the surface of cells [3]. Exosomes are 50 – 100 nm in size and are released via the fusion of endocytic multivesicular bodies with the plasma membrane [3]. Exomeres are the smallest and relatively new subclass of EVs less than 50 nm in size, whose structure lacks a lipid bilayer yet still retain protein markers expressed in microvesicles and exosomes [4]. Certain tetraspanin proteins (CD9, CD63, and CD81) were found to be highly enriched in EVs [5]. CD81 is highly enriched in plasma membranes, whereas CD63 is an endosomal marker [5]. These tetraspanin proteins are also known for their roles in membrane trafficking and oligomerization with other proteins [6]. Depending on their origin, EVs can contain many molecular constituents such as proteins and nucleic acids [7]. Collectively, heterogeneous size, protein makers, and varying composition of EVs add complexity to understanding their roles in biology.

EVs have been reported to play a role in a wide variety of processes including cellular migration, tumor progression, and regulation of immune systems (**Figure 1.2**) [2]. An increasing body of literature has revealed EVs play an important role in cell-to-cell communication [8, 9]. Such exosome-mediated transfer of molecular cargoes have shown involvement with tumorigenesis. Interestingly, Hoshino et al. demonstrated that tumor derived exosomes present certain integrins on their surfaces, which determine organotropic metastasis [10]. Their study revealed that unique integrin combinations preferentially allow the uptake of tumor-derived exosomes into sites of organ-specific metastasis [10]. In addition, oncogenic proteins such as mutant KRAS have been shown to be released into EVs, which enhances the invasiveness of recipient cells [11]. Choi et al. have shown that the oncogene EGFR and its mutant EGFRvIII released from glioblastoma cells are not only present in EVs derived from those cells, but that the proteome of EV-related proteins

changes as well [12]. The role of exosomes in immune response has also been widely documented [13, 14]. Gehrmann et al. revealed exosomes derived from dendritic cells carry NK cell activating ligand which induced both antigen-specific T and B cell activations [13]. Kurywchak et al. discussed how surface proteins on exosomes from B lymphocytes presented major histocompatibility complex (MHC) class I and II contributing to modulation of tumor immunity [14].

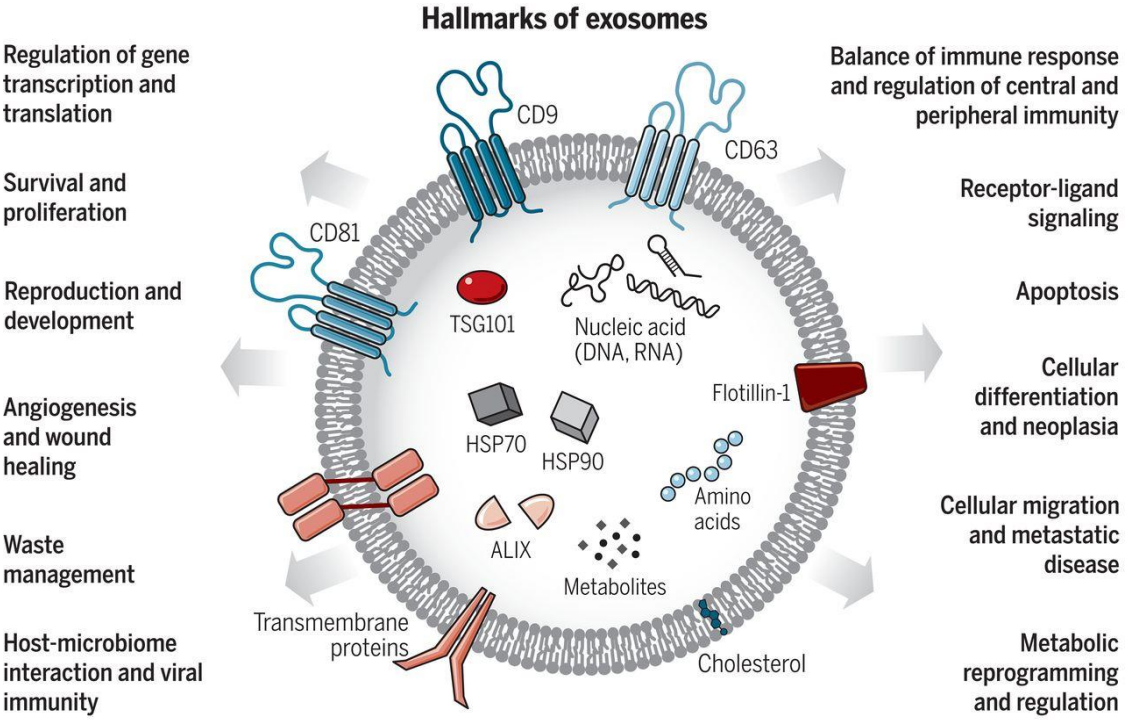


Figure 1.2 | Function and roles of EVs.

The hallmarks of EVs include regulation of gene transcription and translation, survival and proliferation, immune system, cellular migration, and reprogramming. EVs are generated by all cells and carry nucleic acids, proteins, and metabolites, playing important roles in various aspect in human physiology. Reprinted with permission from [2]. Copyright 2020 American Association for the Advancement of Science.

1.3 EV isolation, quantification, and characterization techniques

Despite their promise as circulating biomarkers, EVs have faced key issues in implementation. The foremost concern in utilizing them as liquid biopsy biomarkers is the significant variability in isolation standards, which has brought accompanying pre-analytical variation and further issues with reproducibility [15-18]. This chapter section serves as an overview of the attributes, tradeoffs, and common practices for EV isolation, quantification, and characterization.

1.3.1 EV isolation

Extracellular vesicles can be isolated by a variety of techniques [19]. Various techniques such as ultracentrifugation, ultrafiltration, size-chromatography, polymer-based precipitation, and affinity-based bead capture have been utilized to isolate EVs [19]. Notably, different isolation techniques can affect the size, yield, and ultimately interpretation of results. Compared to cell culture medium, human plasma is a rich source of EVs, other plasma proteins, and lipoprotein particles [20]. Therefore, separation of EVs from plasma proteins and lipoprotein particles have been considerably challenging. As a result, choosing an appropriate isolation technique is critical for the study aims.

1.3.1.1 Ultracentrifugation

Ultracentrifugation (UC) is a conventional method that uses centrifugation force (100,000 – 200,000 x g) to sediment EVs [21]. The efficiency of EV isolation depends on acceleration, type of rotor, and viscosity of the medium [21]. These parameters, which can affect the yield and purity of EVs, should be taken into consideration. Although an

increased centrifugation time can increase the yield of EVs, it can also co-precipitate other plasma proteins [22]. Density gradient centrifugation is another commonly used method which utilizes the inherent density differences in EVs and contaminants such as apolipoproteins in plasma. Density gradient centrifugation typically uses sucrose gradients or commercially available iodixanol gradients (OptiPrep). However, UC-based methods for EV purification are time-consuming and low-throughput, which is not suitable for clinical settings. To overcome these limitations, other simpler isolation methods have been developed.

1.3.1.2 Filtration and size-exclusion chromatography

EVs can be isolated by either molecular weight or size. Ultrafiltration is a method which isolates EVs by the defined size or molecular weight. Typically, molecular weight cut-offs of 10 kDa or filtration < 0.22 μm is expected to concentrate EVs [23]. Although this method is simple, it is difficult to remove contaminating proteins. EV isolation via size-exclusion chromatography (SEC) utilizes porous beads where particles separate by differing sizes. Fractions of solution will be eluted in order of decreasing size, allowing EVs to be separated from smaller plasma proteins. SEC has been shown to provide higher EV purity and good recovery rate [24]. It has advantage in allowing sequential elution of different EV sizes and characterization by transmission electron microscopy which revealed the eluted EVs were intact [25]. Importantly, SEC isolation of EVs has been shown to be rapid and reproducible [26], with single-step plasma EV isolation using SEC published using commercially available Izon qEV SEC columns [27].

1.3.1.3 Polymer precipitation and affinity based bead capture

EVs can also be isolated by their surface properties. Precipitation based methods have relied on reagents which preferentially interact with the phospholipid bilayer exterior of EVs to cause them to aggregate and more easily pellet via centrifugation [21]. The commercial kit ExoQuick™ uses a precipitation reagent mixture which includes Polyethylene Glycol (PEG) polymers [19]. PEG polymers alter the solubility and dispersity of exosomes, facilitating their precipitation from biological fluids [19]. However, co-isolation of soluble proteins and deformation of EVs are unavoidable, affecting down-stream analysis [28]. To overcome the impurities of exosome precipitation, immuno-affinity based methods have been exploited to capture exosomes utilizing specific surface proteins on EV surface membranes. This method utilizes immunomagnetic beads coated with covalently cross-linked streptavidin facilitating biotinylated antibodies against the target molecules. Anti-CD9, -CD63, and -CD81 antibodies are commonly used to isolate exosomes. Based on this technology, a microarray has been developed for exosome detection and phenotyping [29]. While this method is highly specific, it relies on the selected subset of markers which may not reflect all EVs present in a given biofluid. Therefore, this method can be followed by other quantification methods to understand its cargo or remove impurity from the whole plasma.

1.3.1.4 Asymmetric-flow field fractionation

A recent development of size-based fractionation is asymmetric-flow field flow fractionation (AF4) technology. The basic principle behind AF4 utilizes two flows to resolve particles: primary forward channel flow and cross flow perpendicular to the

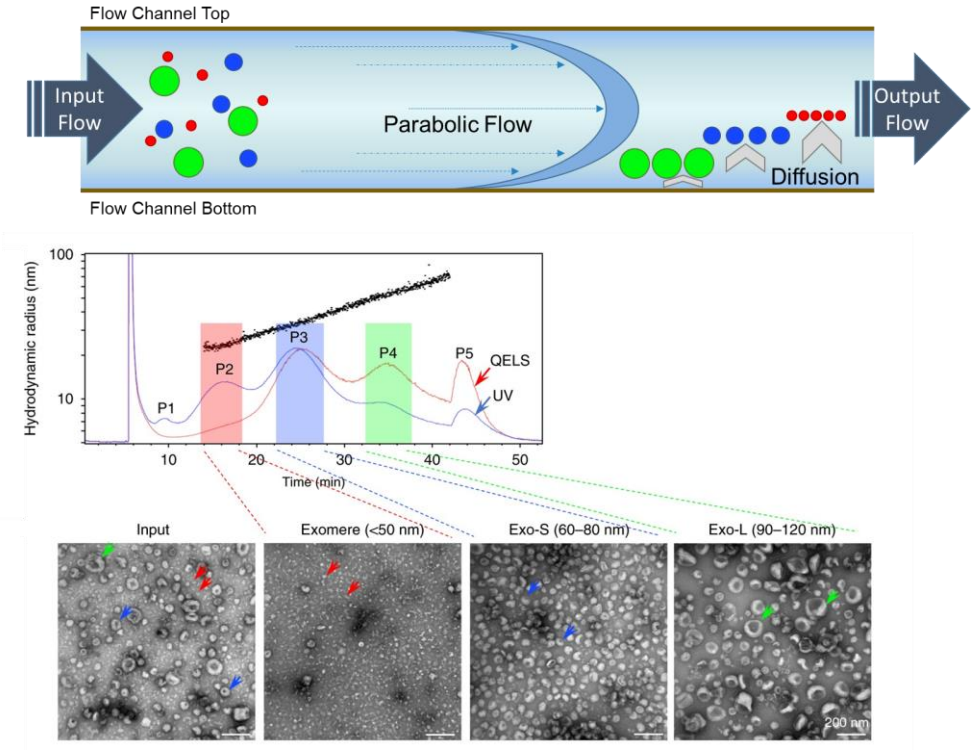


Figure 1.3 | Asymmetric field flow size fractionation of EVs.
Single direction of channel flow from inlet to the outlet is applied, which diffuse the particles with small diameter are eluted at an early time point. Representative AF4 fractionation profiles of B16-F10 derived exosomes with respect to QELS (DLS) intensity (red) and UV absorbance (blue). Transmission electron microscopy images of corresponding fraction revealed heterogeneous EV sizes. Reprinted with permission from [30]. Copyright 2018 Springer Nature.

channel flow [31]. When the sample is applied to thin, flat channel, the input channel flow creates a parabolic laminar flow to move particles from inlet to outlet in a forward direction (**Figure 1.3**). Combined with external physical field by cross-flow, particles are driven toward the bottom of the well. This cross-flow is the driving force to resolve particles with different hydrodynamic sizes counteracting the Brownian motion related to particle sizes [31]. Unlike SEC, AF4 elutes the smallest hydrodynamic sized particles first, as their

higher rate of diffusion lets them most easily stay within the parabolic flow field [31]. In conjunction with real-time monitors such as UV absorbance or dynamic light scattering, distinct size ranges of EVs from cell culture have been separated: so called Exo-L, Exo-S, and exomeres [30]. Although it can separate particles at high resolution in nanometer range, sample dilution through fractionation remains as a major challenge. Collectively, inherent tradeoffs in these different EV isolation strategies must be weighed against specific questions of the study and the impact of isolated EV concentration and purity on downstream analyses. The advantages and disadvantages of most commonly used methods for exosome isolation are summarized (**Table 1.1**).

	UC	MF	PEG	IA	SEC	AF4
Mechanism of separation	Size, density	Size, molecular weight	Surface charge, solubility	Immunoaffinity capture of antigen on surface membrane	Size; large particles eluted first	Size; small particles eluted first
Specificity	++	++	+	+++	++	++
Recovery	++	+	+++	++	+++	+++
Purity	++	+	+	+++	++	++
Time	+	+++	+++	+	++	++

Table 1.1 | Summary of different EV isolation methods.

Specificity which specific exosome isolated is scaled from highest (+++) in immunoaffinity capture to lowest (+) in precipitation method. Recovery which amount of exosomes yields is scaled from highest (+++) to lowest (+). Purity which separates EVs from other contaminants is scaled from highest (+++) to lowest (+). Time for processing is scaled from shortest (+++) to longest (+). UC: ultracentrifugation, MF: membrane filter, PEG: precipitation method, IA: immunoaffinity capture, SEC: size exclusion chromatography, and AF4: asymmetric-flow field flow fractionation.

1.3.2 EV quantification and characterization

Several characterization and validation methods have been developed to analyze EVs. These include: i) biophysical measurements by transmission electron microscopy (TEM), resistive pulse sensing (qNano), and dynamic light scattering (DLS), ii) protein characterization by flow cytometry, western blot, and mass spectroscopy, and iii) understanding of RNA cargo by quantitative polymerase chain reaction (qPCR) and RNA-sequencing.

1.3.2.1 Biophysical property measurement of EVs

Biophysical properties of EVs such as size, shape, surface charge, and concentration are important to understand the basic biology of EVs and their use in applied science. Several techniques have been routinely used to characterize EVs. These include dynamic light scattering (DLS), transmission electron microscopy (TEM) and tunable resistive pulse sensing (TRPS). DLS measures the hydrodynamic particle size distribution resulting from Brownian movement of particles [32]. It is suitable for measuring particles in suspension which are monodispersed [33]. The technique provides diameter range of analyzed particles, however it does not visualize the particles. TEM is widely used to characterize the size, structure, and morphology of EVs [34]. TEM works by focusing the electrons into a very thin beam which is directed to the specimen of interest [34]. Typically, heavy metal stains are used to generate sufficient electron scatter contrast to visualize relatively less dense biological samples of interest [34]. The image formed by the scatter of electrons by the stained sample can be collected either using a fluorescent screen or a charged-coupled device (CCD) [35]. The resulting negatively stained images provide

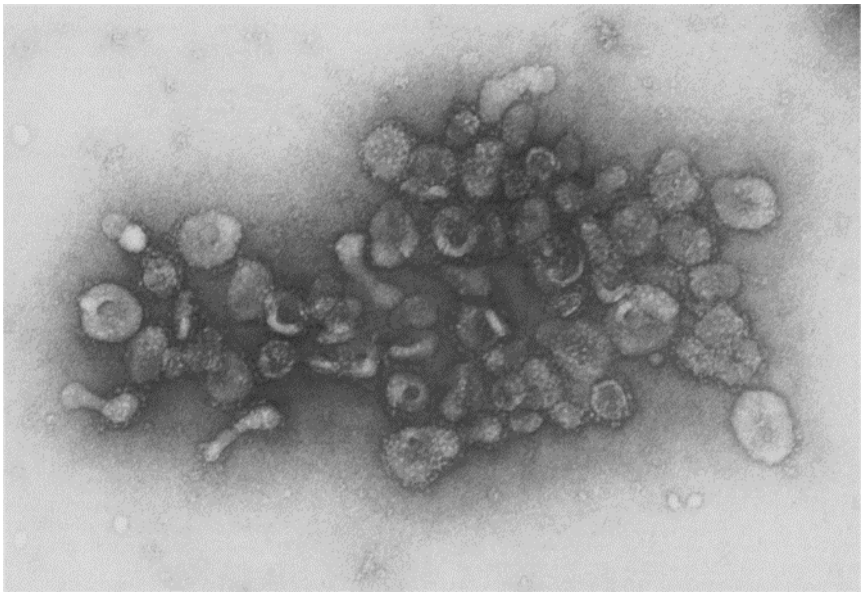


Figure 1.4 | TEM image of transferrin receptor containing extracellular vesicles from sheep reticulocytes.

The first TEM image of extracellular vesicles from sheep reticulocytes. After centrifugation at 12,000 g to remove cells, supernatant was filtered and centrifuged at 100,000 g for 1 hr. The resulting pellet was imaged at 125,000x magnification. Reprinted with permission from [1]. Copyright 1983 Elsevier.

unparalleled detail, often down to nanometers or even angstroms in resolution [34]. The first reported images of extracellular vesicles were in 1983, which used TEM to directly visualize the carrier of radiolabeled transferrin receptors released from sheep reticulocytes over time during cell-culture (**Figure 1.4**) [1]. From its foundations in similar works, TEM has become a gold standard in identifying not only the presence but also morphology of EVs. More recently, advancements in cryo-electron microscopy have enabled the visualization of near-native structures of isolated EVs [36]. Although TEM provides structure, morphology, and size, it can only infer the relative abundance of EVs within a sample. Therefore, alternative methods such as resistive pulse sensing technology are

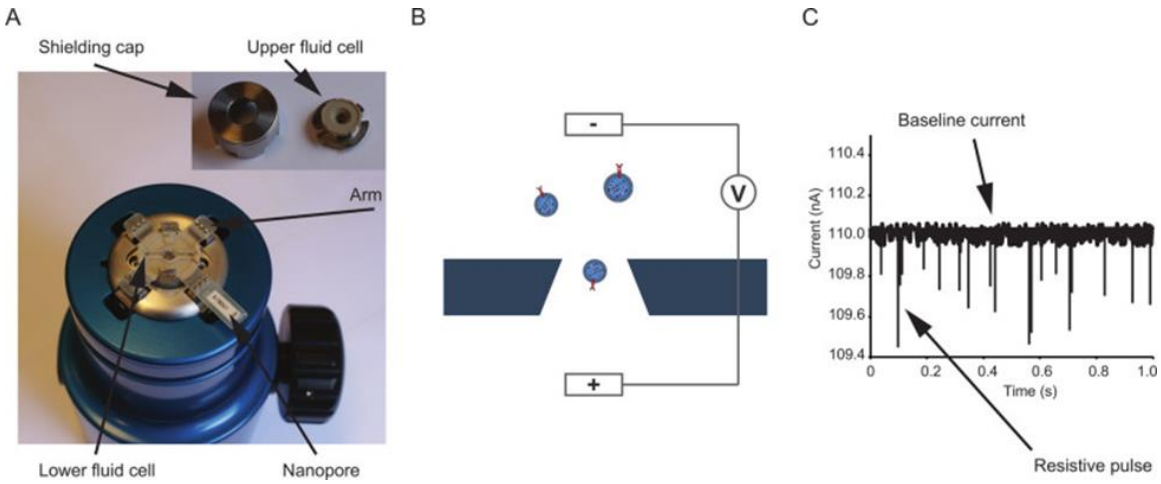


Figure 1.5 | qNano instrument and mode of operation

(A) Commercial setup of resistive pulse sensing with indicated components. Nanopore is positioned on the instrument separating lower fluid shell from upper fluid shell. (B) Schematic diagram of the EVs passing through the nanopore and baseline voltage across it. (C) Example of raw baseline current profile with resistive pulses revealing the detection of passively diffusing EVs. The magnitude is proportional to the size of the particle. Reprinted with permission from [37]. Copyright 2014 Journal of Visualized Experiments.

needed to measure the concentration of EVs. Recently, TRPS has emerged as a new technique. TRPS is a technique which monitors current change when particles pass through a narrow pore (**Figure 1.5**) [38]. Sample particles are driven through the nanopore by applying both pressure and voltage. Each particle causes a resistive pulse or "blockade", and subtle differences in electric impedance are measured over time [38]. The blockade frequency is used to determine the particle concentration, with magnitude directly proportional to the size of the particle [37].

1.3.2.2 Protein Characterization of EVs

Molecular based approaches to characterize EV proteins are essential to understand their biological interactions. Since EVs are lipid bilayer enclosed particles, they include both cell-surface proteins and soluble proteins. EV surface proteins include proteins implicated in antigen presentation, tetraspanins, and lipid anchors [2]. EVs also carry intravesicular proteins such as tumor susceptibility gene 101 (TSG101), heat shock proteins (HSP70 or HSP90), and apoptosis-linked gene 2 interacting protein (ALIX) as internal cargoes [2]. To label and characterize EV proteins, several methods including western blot, flow cytometry, and mass-spectroscopy have been employed. Western blot is a widely used technique to detect the presence of EV-associated proteins in samples of interest. Western blot, however, is often conducted from EVs isolated from cell culture medium where the relative abundance of other contaminants is low. The major challenge in analyzing EVs from complex media such as human plasma is significant contribution of other soluble plasma proteins like albumin [20]. For specific targeting approach, immuno-precipitation methods which utilize magnetic beads coupled with protein A have been developed. The bead-bound antibody permits isolation of EVs with specific targets of interest [39]. However, this method is limited to specific targets and thereby specific subpopulations of EVs. Therefore, other methods such as flow cytometry are increasingly used as a high-throughput and multiparametric technique.

Flow cytometry is currently one of the most popular methods of analyzing EVs [40, 41]. A flow cytometer is a laser-based instrument for analyzing physical characteristics of cells or particles of interest. It is commonly used to analyze relative size, relative internal complexity, and relative fluorescence of labeled antigens on the surface of cells [42]. Flow

cytometers are composed of three main components (fluidics, optics, and electronics) to work simultaneously together for particle detection and analysis. The fluidic system focuses particles to laser beam for interrogation. Typically, samples are injected into the center of a pressurized buffer stream (sheath fluid). The pressurized sheath fluid is driven through the illumination path, forcing the sample pass through a flow cell [42, 43]. Sheath fluid density and velocity differ from the sample, creating a laminar flow which does not mix with the sample [43]. This confines a slow flowing stream by the faster flowing stream, known as hydrodynamic focusing. The hydrodynamic focusing guides the sample particles in a single-file stream, wherein a series of lasers are focused onto the sample at the interrogation point [43]. The optical subsystem then provides the excitation sources and detector components. A series of lasers and an array of filters in front of the detectors (typically photomultiplier tubes or photodiodes) are used to detect and parse the different wavelength emissions of common fluorophores which are used for the immunodetection of the particles of interest. Finally, the electronic subsystem converts light signal to electronic signals, providing numerical values for pulse height, width, and area [43].

Light scattering occurs when the particle passes through the laser beam. Forward scattering is detected in the forward direction of the laser beam. Side scattering is detected approximately 90° to the laser beam. Light scattering is affected by both size, refractive index, and granularity of the particle in a fluid. Determination of accurate light scattering and size have been challenging as most EVs cannot be resolved by light microscopy and their diameter is notably smaller than visible light wavelengths [44]. Recently, the refractive index of individual small EVs have been assessed utilizing nanoparticle tracking analysis [44, 45]. The relative light scattering intensity values of the individual particles

were compared to theoretical light scattering generated from Mie theory to solve the unknown refractive index of particles. A laser beam with specific power illuminates each particle in suspension generating scattered light. The trajectory of each particle moving by Brownian motion is used to determine the diffusion coefficient, which can be related to the particle diameter via Stokes-Einstein equation [46, 47]. Mie theory is used to derive relationship between scattered light and refractive index [45]. The vesicle consists of a several nm thick phospholipid shell resulted in a refractive index of 1.46 ± 0.06 [45].

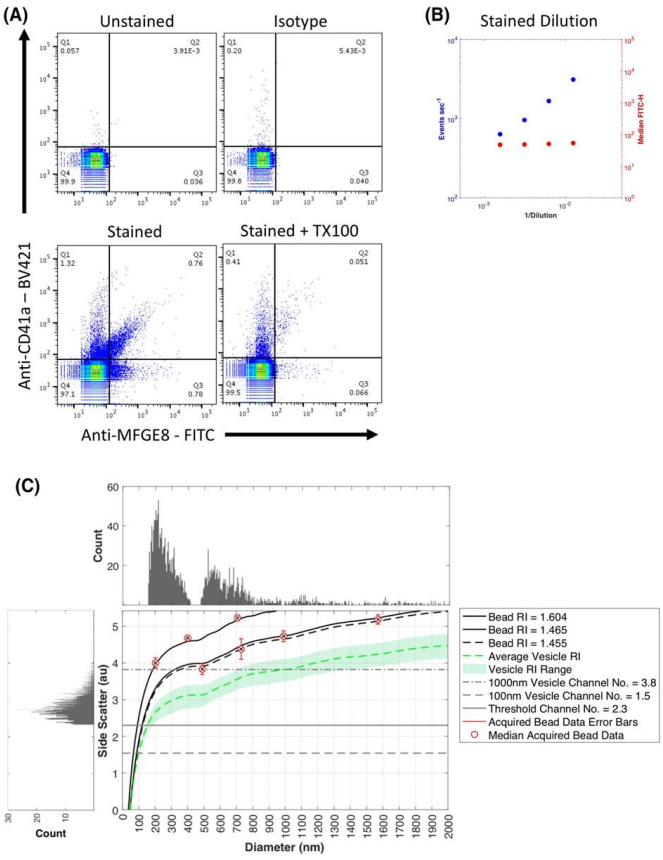


Figure 1.6 | Analysis of platelet-derived EVs using flow cytometer

(A) Dot plot of unstained, isotype control, CD41 and phosphatidylserine stained EV, and detergent treated stained EV. (B) Stained EV dilution control with consistent median FITC-H intensity. (C) Scatter-diameter curve relationship from FCMpass software. Acquired bead data are shown with predicted model and green region indicates vesicle diameter with effective refractive indices. Reprinted with permission from [40]. Copyright 2019 Wiley.

Utilizing the refractive index of EVs, a growing body literature describes importance of light scatter standardization to facilitate highly consistent and reproducible data between commercial flow cytometers. Calibration of light scatter is demonstrated by Fattacioli et al., and specifically for EVs by van del Pol [41, 48]. Welsh et al. developed the software (FCMpass) package for light scatter and fluorescence [40, 49, 50]. The software composed of fluorescence calibration and light scatter calibration for standardized EV reporting units [40, 49, 50]. Fluorescence calibration is performed by using molecules of equivalent soluble fluorochrome (MESF) beads. These are fluorescent microspheres that are labelled with specific amounts of fluorophores. The MESF units are determined by comparing fluorescence intensity signal from the microbeads standards to the signal from a solution of the same fluorochrome [51]. Using the assigned MESF unit, fluorescence intensity can be standardized between varying instrument sensitivity. Similarly, light scattering calibration can be done using reference beads from the National Institutes of Standards and Technology (NIST). Mie modeling and subsequent conversion of observed light scatter intensity to diameter can be performed using FCMpass software (**Figure 1.6**) [40, 49, 50].

Mass spectrometry has also been used to screen for proteomic profiles of EVs. In this approach, proteins extracted are digested into peptides that can be subsequently separated by liquid chromatography and analyzed by the mass spectrometer [24]. Relative quantification of protein abundance is useful to employ EV proteins as novel biomarker sources. During the last decade, high-resolution mass spectrometry (HR-MS) has become established. HR-MS utilizes tandem mass spectra coupled to nano-HPLC. There are two approaches to make MS quantitative: isotope-based [52] and label-free method [53].

Absolute quantification is determined by comparing ion intensity between isotope-labeled authentic standards and analyte as physical properties of isotope labeled compounds are identical [54]. However, limited availability of isotope standards restricts the number of samples which can be accurately compared. Therefore, advancements in label-free methods have enabled quantification in a large scale without additional experiment steps [54]. These proteomic approaches have been used in several EV characterization studies. For example, proteomic studies of EVs have revealed heterogeneous populations of EV subtypes [55]. EVs from human primary monocyte-derived dendritic cells were first separated by differential centrifugation followed by either iodixanol density gradient or by immuno-isolations [55]. Comprehensive proteomic analysis of different EV isolation methods yielded histocompatibility complex, flotillin, and heat shock protein present in all EVs [55]. For studying EV proteins in relation to cancer, Beckler et al. purified exosomes from two colon cancer cell lines. They confirmed EV-sized particles through TEM and positive presence of EV-enriched markers HSP70, FLOT1, and TSG101 via western blotting [11]. Intriguingly, their mass spectroscopy analysis from exosomes derived from colon cancer cell lines with KRAS mutant allele (DKO-1) and KRAS wild-type allele (DKs-8) revealed that mutant KRAS caused an increase in proteins related to vesicle components or transport (16% of upregulated) while losing RNA-binding associated proteins (30% of downregulated) [11]. Further advances have utilized human plasma from multiple cancers [56]. Using AF4, proteomic profiling of EVs revealed pan-EV markers from cells, tissues, and most biofluids in human and murine samples: CD9, HSPA8, ALIX, HSP90AB1, and ALIX [56]. Compared to tumor tissue explant and normal tissue, Zhang et al identified tumor differentially expressed proteins (VCAN, TNC, and THB2) with high

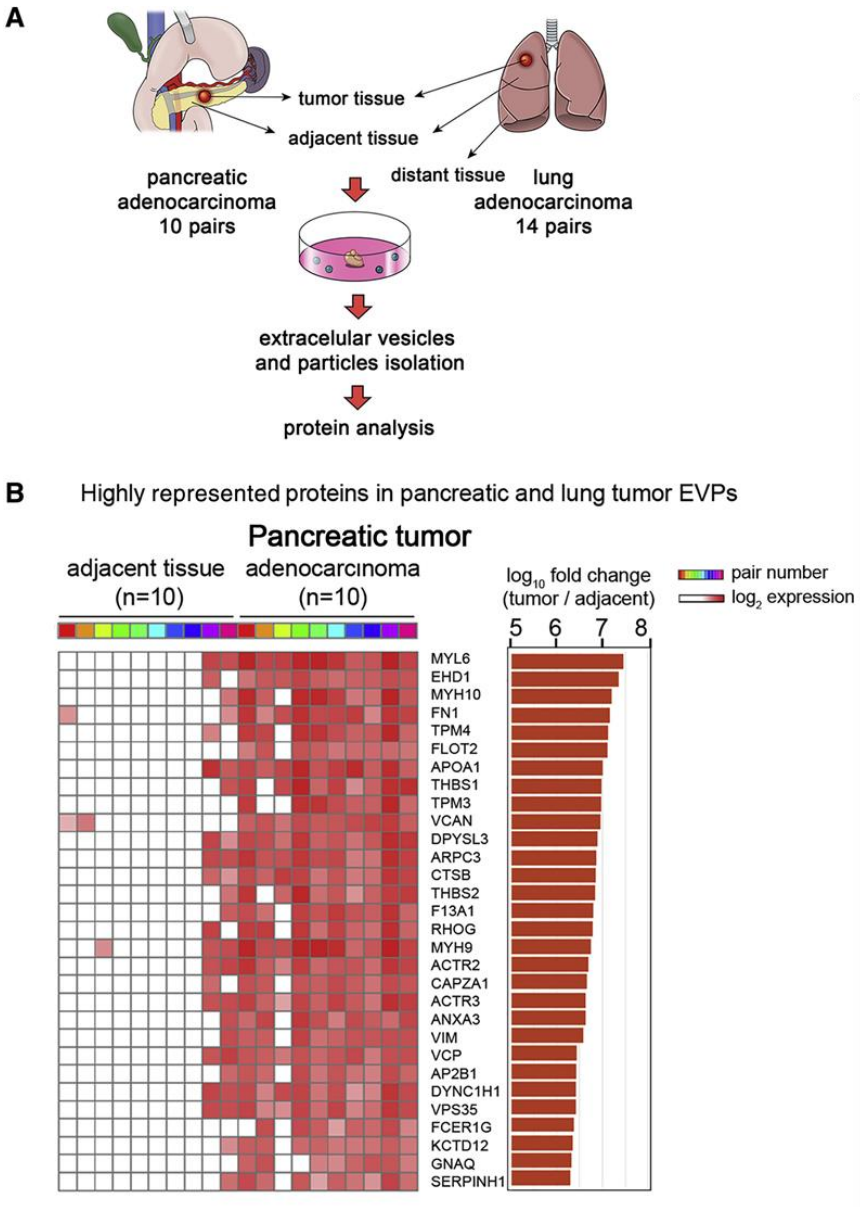


Figure 1.7 | Proteomic analysis of EVs from surgically removed pancreatic cancer tissue explant

(A) Diagram of tissue explant method from tumor tissue, adjacent tissue, and matched distant tissue. Millimeter-sized fresh tumor and peritumoral adjacent tissue were harvested from patients with localized pancreatic cancer or lung adenocarcinoma. Tissue was cut into small pieces and cultured for 24 hours in serum-free RPMI. Conditioned media was processed for EV isolation. (B) Top 30 proteins highly represented in pancreatic tumor tissue compared to adjacent tumor tissue. Reprinted with permission from [56]. Copyright 2020 Elsevier.

sensitivity and specificity (**Figure 1.7**). Importantly, Zhang et al. performed proteomic profiling on plasma and found 51 and 19 unique proteins to pancreatic cancer and lung cancer [56]. Taken together, plasma-derived EV protein profiles from various sources could serve as liquid biopsy tools to detect cancer.

1.3.2.3 RNA characterization of EVs

In addition to proteomics based approaches, the assessment of coding and noncoding RNAs in EV is undergoing intense research. RNA profiling has emerged as a powerful tool to investigate the potential of EV derived RNA as a biomarker from human biofluids. However, RNA sequencing from biofluids is technically challenging due to their low input amount and the degradation of RNA [57]. To establish whether extracellular RNA and their carriers such as extracellular vesicles may mediate intercellular communication, the Extracellular RNA Communication Consortium (ERCC) was launched by NIH Common Fund to establish foundational knowledge about extracellular RNA research [58]. Extracellular vesicles have been shown to play an important role in transporting RNAs between cells and promote tumor growth [59]. How these RNAs reside in RNase-rich biofluids, specific types of cargoes associated with subtypes of EVs, and overall clinical utility remain to be established [58]. Importantly, efforts to develop technologies, informatics tools, and analysis are essential in related research studies.

Using computational deconvolution, the data repository of extracellular RNA communication consortium analyzed cell free RNA cargo from various biofluids covering 23 health conditions across 19 different studies (**Figure 1.8**) [60]. Their computational integrated analysis revealed within the non-coding RNAs, there are 4 major types of

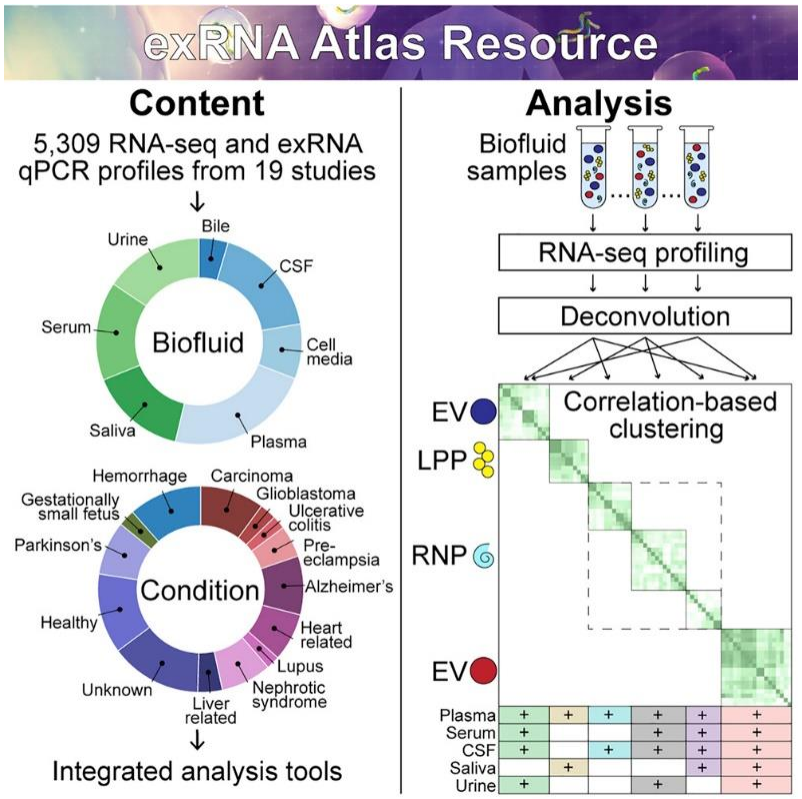


Figure 1.8 | ExRNA atlas from Extracellular RNA Communication Consortium

The NIH Extracellular RNA communication Consortium created exRNA atlas resources containing 5,309 exRNA sequencing and exRNA qPCR profiles using 7 different bodily fluid from 19 different studies. To analyze variation between studies, computational devolution analysis was used to model six cargo types for non-coding RNA. Reprinted with permission from [60]. Copyright 2019 Elsevier.

carriers: high and low density EVs, lipoprotein particles, and RNA binding proteins [60]. Due to substantial overlap in their physicochemical properties, identifying effective separation and characterization of these known carriers has remained a challenge. EVs and high-density lipoproteins have similar density which can co-fractionate using density-gradient ultracentrifugation [61]. However, size-exclusion chromatography has been shown to separate these two particles based on size [58], although size resolution within

the EV peak to identify EV heterogeneity remains challenging. Importantly, sequencing of small RNA using size-exclusion chromatography revealed RNA are sorted into EVs or RNA binding proteins [62]. Relevant to the packaging of RNA cargo into carriers, another study revealed how oncogenes, such as KRAS, may influence the selective packaging of genetic materials into vesicles in cell culture media [63]. However, little is known for cell free messenger RNA, especially in human biofluids, regarding the major type of carrier and understanding how cancer may dysregulate RNA packaging profiles.

1.4 Fundamentals of cell-free RNA

Cell free RNA (cf-RNA) is another major group of circulating biomarkers and holds promise in disease detection and diagnosis. Circulating RNAs are highly specific and amplifiable which makes them an ideal target as novel tools in cancer diagnosis. The first discovery of circulating nucleic acids originated back in 1948 as described by Mandel and Metais [64]. Surprisingly, it wasn't until 1999 that cell-free RNA (cf-RNA) was first discovered, when two groups identified circulating messenger RNA (mRNA) in the plasma of patients with Nasopharengeal carcinoma and malignant melanoma. Lo et al. found cell-free Epstein-Barr viral mRNA in the plasma of patients with nasopharengeal carcinoma [65]. Additionally, despite generally higher serum RNAase activity in patients with malignant melanoma, Kopreski et al. found elevated Tyrosinase mRNA in patient serum which passed through a 0.45μm filter [66]. Since then, a wide range of RNAs have presented in human biofluids: messenger RNAs (mRNAs), long noncoding RNAs (lncRNA), microRNAs (miRNAs), circular RNAs (circRNAs), tRNA-derived fragments (tRFRNAs) and Piwi-interacting RNAs (piRNAs) [67].

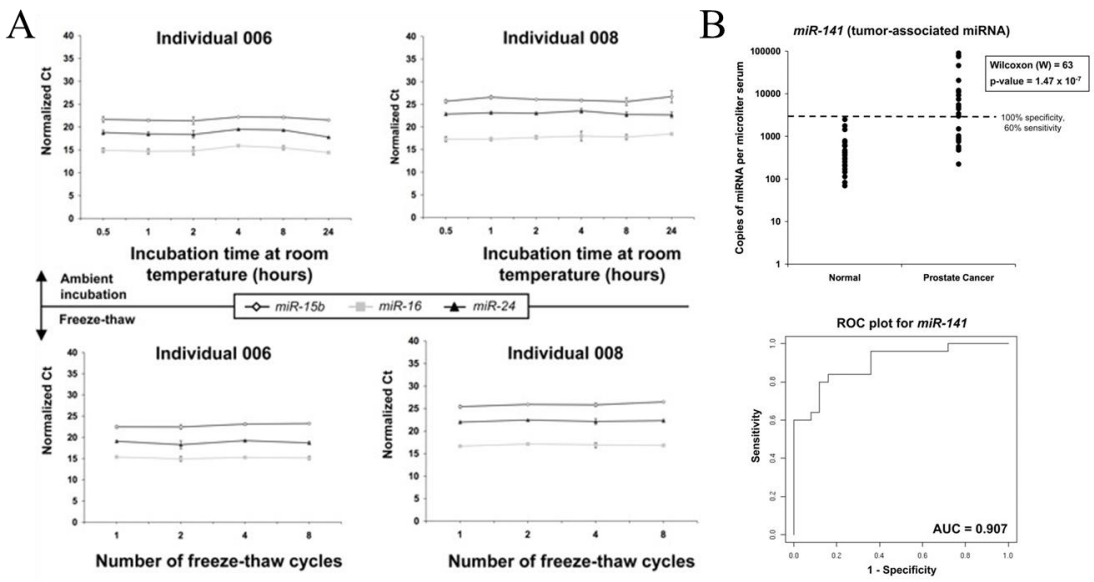


Figure 1.9 | Characterization of miRNA stability and detection of human prostate cancer by serum levels

(A) Normalized count threshold (Ct) results for indicated miRNA subjected to prolonged room temperature incubation or multiple freeze-thaw cycles. (B) Detection of human prostate cancer serum level of tumor associated miR-141. Receiver operation characteristic (ROC) plot for miR-141, showing area under the curve value of 0.907. Reprinted with permission from [68]. Copyright 2008 National Academy of Sciences.

To accelerate the progress in the new field of cell free RNA biology, the Extracellular RNA Communication Consortium was launched in 2013 [58]. The consortium is designed to overcome many gaps in knowledge and technical challenges. These include, but are not limited to: i) understanding the biology of cell free RNA, ii) biomarker discovery in human diseases, and iii) development of technical tools including bioinformatics. Mitchell et al. investigated the detection and stability of miRNA in human plasma (**Figure 1.9**). microRNAs (miRNAs) are approximately 21-22 nucleotide non-coding RNA molecules that regulate gene expression at the post-transcriptional level [68,

69]. Despite the exogeneous RNAse activity in human plasma, levels of miRNA were stably detected over 24 hours of incubation and after up to eight cycles of freeze-thawing [68]. The mechanism of this protection is a growing current research topic which can be associated with either EVs, ribonucleoprotein or lipoprotein complexes. In regards to biomarker discovery in human cancer, Mitchell et al. revealed miRNA-141 levels were overexpressed in patients with prostate cancer from healthy control [68]. Sayeed et al. investigated cell-free mRNA transcriptome in people with liver cirrhosis (LC) and hepatocellular carcinoma (HCC) to determine biomarker potential of cell-free mRNA [70]. Using RNAseq and RT-qPCR, liver-derived circulating transcripts were significantly upregulated in HCC patient samples revealing potentials for cancer detection [70]. Despite initial findings in these studies, whether transcripts can differentiate pan-cancer remains unknown. To develop a computational analysis tool to understand different miRNA cargo types and its association with cancer, Extracellular RNA Communication Consortium created a data repository between studies and developed bioinformatics tools [60].

1.5 RNA isolation, quantification and characterization techniques

The human circulation contains cell free RNA, which can be an important source of biomarkers. An earlier study demonstrated miRNA are released into the circulation in a remarkably stable form after longer duration of incubation time, and even after many freeze-thaw cycles [68]. Additionally, this work revealed circulating miRNA carry disease specific signatures that can be exploited as non-invasive biomarkers. Despite their promise as circulating biomarkers, there is a critical need to provide more reliable and reproducible

results from human biofluids. In this section, sample and assay standards focused on standardization of RNA isolation and profiling methods will be discussed.

1.5.1 RNA isolation

1.5.1.1 RNA extraction

One of the most exciting areas of cell free RNA research involves the assessment of cell free RNA present in serum or plasma samples. For the purification of cell-free RNA, especially miRNA, there are many different commercial kits: i.e. RNAdvance (Agencourt Bioscience, Beckman Coulter, Beverly, MA), MAgMAX (Life Technologies, Thermo Fisher Scientific), miRCURY-Biofluids (Exiqon, Vedbaek, Denmark), Quick-RNA (Zymo Research, Irvine, CA), DirectZol (Zymo Research, Irvine, CA), miRNeasy (Qiagen, Hilden, Germany), and mirVana (Thermo Fisher Scientific) [71]. Most commonly used miRNA isolation kit utilize: 1) TRIzol (guanidium-acid-phenol extraction) reagent, 2) proprietary paramagnetic, or/and 3) silica bead-based technology. The organic extraction technique using TRIzol is widely used in molecular biology for RNA isolation [72]. The single step technique was originally published by Piotr Chomczynski and Nicolette Sacchi in 1987 [73]. This method relies on phase separation from a mixture of an aqueous sample and solution containing phenol and chloroform. Guanidium thiocyanate, a chaotropic agent, is added to organic phase to denature proteins which bind nucleic acids [74]. Under acidic pH, nucleic acids partition into different phases allowing DNA and RNA to be separated due to differences in protonation of DNA and RNA [74]. Precipitation of nucleic acids is then performed using ethanol while the resulting pellet is resuspended in Tris-EDTA (TE) buffer [74].

Solid-phase nucleic acid was originally developed to avoid toxic phenol and chloroform phase separation. Solid-phase nucleic acid extraction allows quick and efficient purification compared to conventional methods [74]. The silica or paramagnetic beads absorb nucleic acids during the extraction process relying on the pH and salt concentration of the buffer. For magnetic beads, biopolymer such as cellulose which exhibits affinity to target nucleic acids are modified into the surface [74, 75]. However, this approach is not nucleic-acid specific and can also absorb other biosubstances as a drawback [75]. To help alleviate this, silica based isolation has been developed utilizing specific charge interaction [76]. In general, negatively charged nucleic acids bind tightly to silica particles under high ionic strength ($pH < 7$) and can be eluted under low ionic strength ($pH \geq 7$) [74, 76]. Silica carbide (SiC) based DNA/RNA isolation also avoids toxic phenol and chloroform phase separation while retaining a wide variety of nucleic acid lengths and content. When comparing to other total RNA isolation kits utilizing either silica fiber (SiF) or silica membrane (SiM), SiC based methods have the highest total recovery of RNA, especially including low molecular weight RNA such as miRNA [77] (**Figure 1.10**). Silica based columns are typically engineered to retain either high or low molecular weight RNA, while SiC beads retain the full complement of smaller miRNA and larger RNA molecules [77]. Therefore, optimization of RNA isolation steps is a critical aspect of study design.

The choice of different RNA isolation kits relies on the type of biofluid and subtypes of RNA being studied [78]. Comparison of different commercially available RNA extraction kits has been evaluated from plasma [78]. miRNeasy serum/plasma and miRNeasy serum/plasma advanced both extracted total RNA while being enriched for small RNA populations < 200 nt [78]. Quick-cfRNA serum and plasma is targeted for both

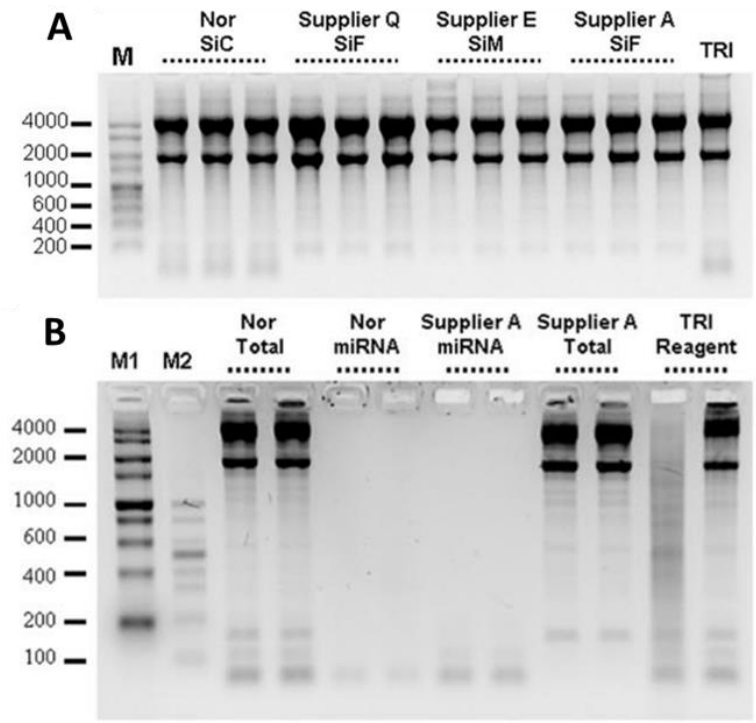


Figure 1.10 | Total RNA isolated using silica carbide and other RNA extraction kits
(A) Performance of silica carbide in RNA purification was compared to the current technology using phenol-based and silica-based extraction in 1.5% formaldehyde-agarose gel. Total RNA was isolated from HeLa cells employing either silicon carbide (SiC), silica fiber (SiF) or silica matrix (SiM). Guanidine thiocyanate/phenol based TRI reagent isolation was used as a positive control for complete size range of RNA isolation. Only SiC and TRI contained both the large and small RNA species. (B) Recovery of small RNA was compared to two commercially available miRNA kits. Enrichment of small RNA using SiC did not involve any phenol extraction, but enriched miRNA similar to those with phenol extraction. Reprinted with permission from [77]. Copyright Norgen Biotek Corp. Thorold, ON, Canada.

miRNA and mRNA, and Isolate II fractionates RNA population based on size to select for small RNAs [78]. Srinivasan et al. presented 10 different extracellular RNA isolation methods across 5 biofluids using small RNA sequencing [79]. Importantly, RNA size distribution and yield varied according to the biofluid type and extraction method used

(**Figure 1.11**). Using small RNA-sequencing, Srinivasan et al. identified the distribution of RNA biotypes also varied by different RNA extraction kits, and complexity was correlated with differences in read depth. Therefore, RNA isolation methods should be selected based on RNA types of interest to observe meaningful results within a given study.

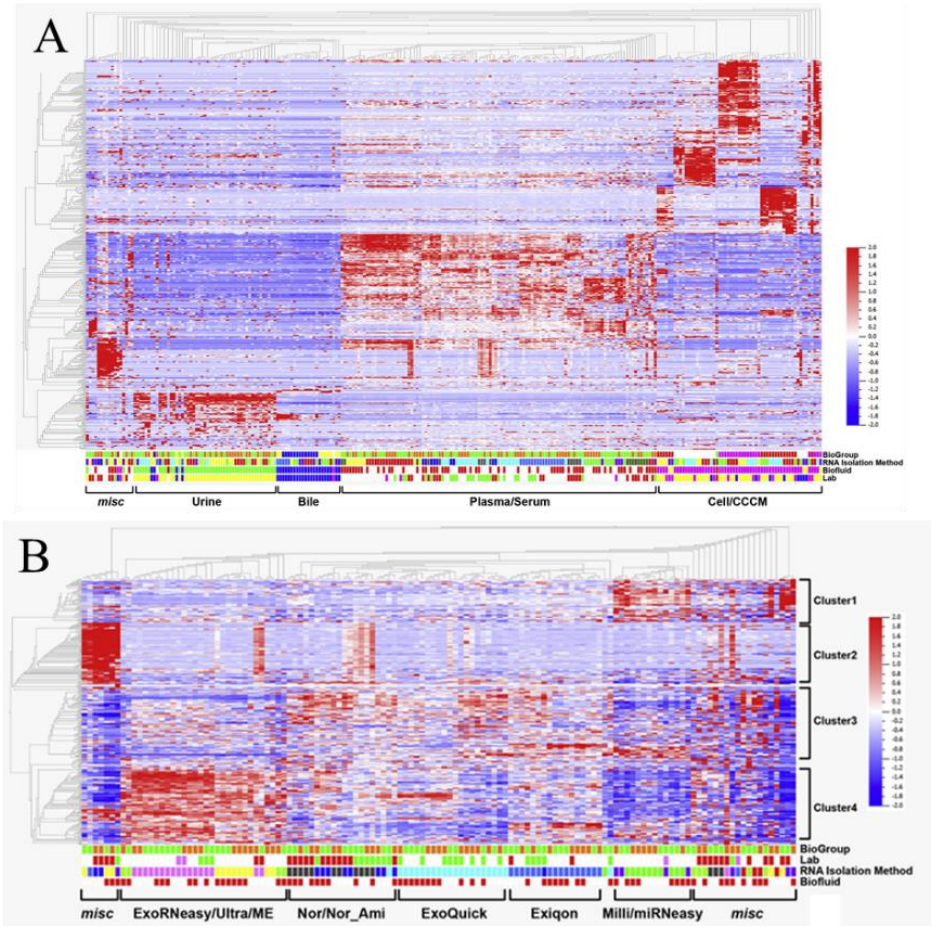


Figure 1.11 | Hierarchical clustering analysis of miRNA among different biofluid types and exRNA isolation methods

(A) Heatmap of miRNA unsupervised clustering of samples by biofluid types (bile, plasma, serum, urine, and cell culture medium). (B) Heatmap of miRNA within plasma and serum samples by exRNA isolation methods. Colors shown for biological group, biofluid, lab and exRNA isolation method. Reprinted with permission from [79]. Copyright 2019 Elsevier.

1.5.2 RNA quantification

Accurate determination of the RNA yield is important for downstream applications since qPCR and next generation sequencing (NGS) require specific concentrations for optimal performance. RNA quantification can be performed using the following methods: i) UV absorbance, ii) fluorescence measurement using nucleic acid dye, and iii) qPCR.

1.5.2.1 RNA quantification by UV absorbance

The most common technique used to determine RNA concentration and purity is UV absorbance. Absorbance is used to measure total RNA concentration in a purified sample. Absorbance of 260 nm light relative to blank buffer approximates both RNA concentration and purity [80]. Absorbance at 260 nm light (A260) is used as nucleic acid bases in the RNA molecule most strongly absorb light at this wavelength [80]. The RNA concentration is calculated using the Beer-Lambert law, $A = \epsilon Cl$, where ϵ is the extinction coefficient (ϵ for RNA is 0.025 $(\mu g/ml)^{-1}cm^{-1}$, C is the concentration of the nucleic acid, l is the path length of the cuvette, and A is the measured absorbance at 260 nm [81]. To evaluate protein contamination, the ratio of absorbance at 260 nm (nucleic acid absorbance) and 280 nm (absorbance of peptide bonds) is used [81]. Typically, A260/A280 ratios over 1.8 are considered highly pure RNA [81]. For RNA measurement, the presence of DNA would falsely indicate higher RNA abundance as DNA contaminants share the same base pair absorption at 260 nm [81]. To avoid this, samples can be treated with an enzyme called deoxyribonuclease (DNAse) which specifically digests DNA and not RNA. The A260/A280 ratio is specifically affected by buffer pH, where studies have shown that weakly basic conditions enhanced the sensitivity of this ratio in determining nucleic acid

purity [82]. For samples with very low RNA concentrations, UV spectroscopy is not sensitive below 0.1 at 260 nm absorbance, or 4 µg/ml RNA [83]. Although spectroscopy is commonly used, its sensitivity and specificity to distinguish DNA, RNA, or protein can be unreliable and inaccurate [84, 85].

1.5.2.2 Fluorescence measurement using nucleic acid dye

In light of drawbacks for quantifying RNA using UV spectroscopy, fluorescence based measurements have become a common alternative [86]. Two of the most commonly used fluorescence-based RNA quantification systems are Qubit fluorometer and bioanalyzer. The most significant difference between fluorescence-based quantification and UV absorbance is the specificity of the molecules of interest (RNA, DNA, or protein). Fluorescence-based methods leverage distinctive fluorogenic dyes which can exhibit > 200-fold enhancement for binding RNA [87]. The Qubit fluorometer can detect 250 pg/µl to 100 ng/µl, however, there is no information about the size distribution [88]. The bioanalyzer system, which can detect as little as 50 pg of total RNA, can provide both the size and abundance of RNA [87, 89]. The bioanalyzer works using a microfluidics chip incorporating both gel and nucleic acid intercalating dyes. Similar to RiboGreen^{TM}, the RNA Nano Dye used by the bioanalyzer interacts with single-stranded RNA molecules to sensitively permit fluorescence versus unbound fluorophores [90]. Another key feature of the bioanalyzer is the electropherogram. As samples move through microchannels, samples are electrophoretically separated [87]. Smaller size RNA migrate faster through the microchannel than the larger ones [87]. The fluorescent signal from different RNA lengths are measured into gel like images (bands) and electropherograms (peaks) [87]. The gel-like image produced from the bioanalyzer is similar to a standard agarose DNA/RNA gel

stained with ethidium bromide. [87]. Quantification is done using RNA ladder as a reference. Overall, the bioanalyzer provides better sensitivity towards low input RNA and also provides measurements of RNA integrity and sample purity.

1.5.2.3 qPCR mechanism and detection

Polymerase chain reaction (PCR) has become a central technique in biochemistry and molecular biology for RNA quantification. PCR was invented in 1983 by Kary Mullis, awarded the Nobel Prize for the procedure to replicate DNA [91]. Given a small amount of RNA, two nucleotides, DNA polymerase, and four deoxynucleoside triphosphate (dNTPs), millions to billions of copies of specific DNA can be generated [92]. Compared to aforementioned quantification methods, qPCR measures the concentration in real time during the exponential phase of the amplification product [93]. PCR relies on a thermal cycling process which goes through different phases: 1) denaturation melts DNA double helix at high temperatures (94–98 °C), 2) annealing steps in which primers bind to complementary sequences of template DNA at lower temperature (55–70 °C), and 3) extension and elongation which DNA polymerase enzymatically assembles new DNA strands using dNTPs [94, 95]. Reverse transcription-qPCR (RT-qPCR) is a technique which reverse transcription or the RNA template into DNA occurs prior to PCR to amplify complementary DNA (cDNA). As the PCR enzyme strictly recognizes double stranded DNA (dsDNA), RNA samples must first be converted to DNA using reverse transcriptase (RT). RT uses a RNA as a template to synthesize complementary DNA (cDNA). Typical number of cycles is usually carried out 25-35 times depending on input amount for the desired yield of the PCR product [96]. Performing more than 45 cycles is not recommended as nonspecific bands start to appear.

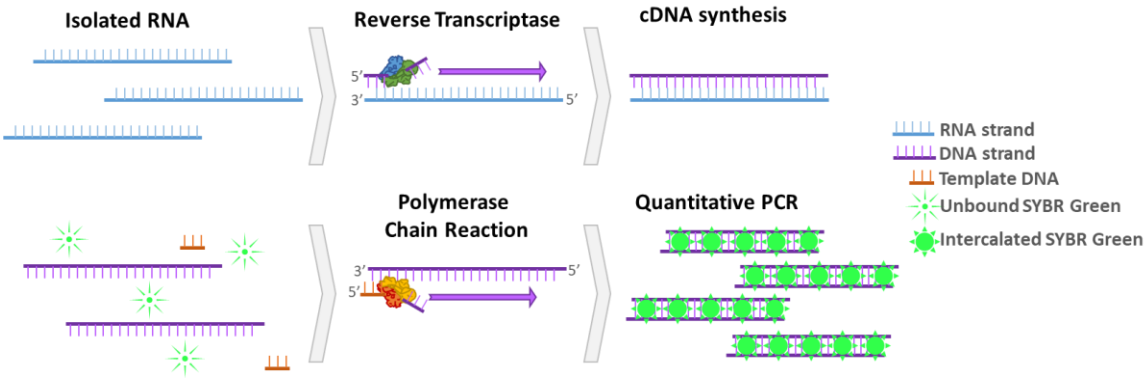


Figure 1.12 | Overview of qPCR workflow for measuring RNA using SYBR green. Reverse transcriptase converts purified RNA into complimentary DNA strands. Polymerase chain reaction, with gene-specific template DNA and SYBR green dye, enables real-time quantification of PCR amplification and relative gene abundance.

During each PCR cycle, the presence of DNA-intercalating dyes enables a fluorescent readout of DNA concentrations (**Figure 1.12**). There are two methods for simultaneous detection and quantification. One is using fluorescent dyes that are retained nonspecifically between double strands. The other one involves probes which specifically bind target sequences to become fluorescently labeled. SYBR green functions as an intercalating dye, whereby segments of the cyanine-based dye will insert into the pi-orbital stacks between nucleic acid bases of dsDNA [97]. During each PCR cycle, SYBR green dye binds to double stranded products resulting in a net increase of fluorescence [98]. However, SYBR green will nonspecifically bind all dsDNA, requiring individual targets to be amplified in separate reaction wells. DNA-purity can be inferred by performing a melt-curve analysis and measuring the dissociation of SYBR dye. The melting temperature of the specific amplified product should yield a sharp decrease in fluorescence when the target dsDNA melts into single stranded DNA (ssDNA) and SYBR green intercalating

fluorescence becomes inhibited [97]. On the other hand, Taqman probe-based assay is specific to the target of interest using fluorescence resonance energy transfer (FRET) (**Figure 1.13**). Base oligonucleotide template sequences are made to match only unique DNA sequences for specific genes with one side conjugated with fluorophore and quencher on the other side. When each fluorophore is intact with a quencher, the proximity of the quencher prevents fluorescence emitted by the fluorophore. Within each PCR cycle, the complimentary probes will be paired and cleaved from 5' end by DNA polymerase, releasing originally bound quencher and enabling fluorescent readout [99].

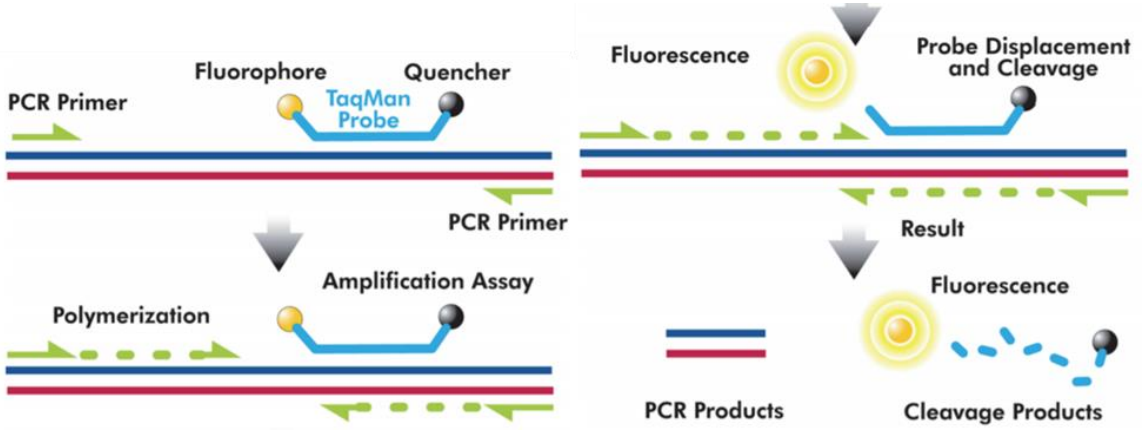


Figure 1.13 | Overview of RT-qPCR workflow for measuring RNA using TaqMan. TaqMan probe relies on 5'-3' exonuclease activity of Taq polymerase to cleave dual-labeled probe during hybridization. TaqMan probe has fluorophore on one side and quencher on the other side, which quenches the fluorescence emitted by fluorophore. Degradation of the probe releases the fluorophore, which can be detected in qPCR. Reprinted with permission from [100]. Copyright Agilent Technologies, U.S.A.

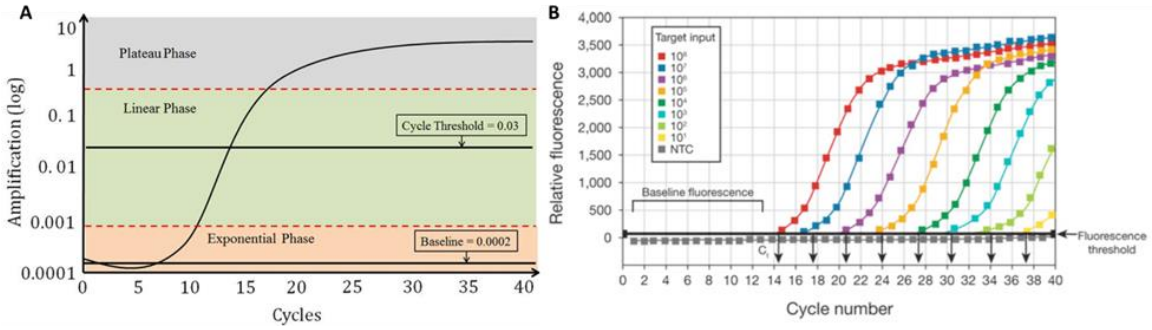


Figure 1.14 | Basic principle of PCR and relative fluorescence cycle number
(A) Baseline, exponential phase, linear phase, and plateau phase during PCR amplification. Reproduced with permission from [101]. Copyright Rice et al. (B) Amplification plots are created when fluorescent signal is plotted for every cycle number. The higher concentration of target input results in lower cycle threshold value. Reprinted with permission from [102]. Copyright Life Technologies Corporation, Carlsbad, CA, U.S.A.

When calculating results of a quantification assay, either relative or absolute quantification is used. There are three phases of fluorescent intensity in a RT-qPCR plot: 1) The exponential phase, where input cDNA is amplifying exponentially yet fluorescent readout is not above background fluorescence, 2) the linear phase, where cycle causes a linear observed fluorescence increase on a log scale, and 3) the plateau phase, where PCR products are so abundant that fluorescent probes saturate the detector nonlinearly (**Figure 1.14**). The parameter, cycle threshold, is defined as the number of cycles at which the fluorescence passes the fixed threshold. This threshold is defined from an average standard deviation of fluorescence emission intensity of the reporter dye [97]. In general, relative quantification is used to analyze changes in gene expression between treatment and control groups. The cycle threshold (Ct) value is inherently relative and it relies on: reagent quality, PCR efficiency and instrument calibration. However, when a reference DNA with known concentration is also added, the relative nature of Ct values then enables absolute

quantification. By interpolating their quantity from a standard curve, the absolute copy number of specific DNA can be obtained.

1.5.3 RNA sequencing, alignment, and characterization

Following RNA extraction and quantification, RNA can be sequenced to enable transcriptomic analysis and comparative gene expression from different diseases and treatments. RNA sequencing is a technique which examines the quantity and sequences of RNA in a sample using next generation sequencing (NGS) technology. NGS is a powerful technology revolutionizing total transcriptomic profiling to understand expression levels for both coding and non-coding RNAs [39, 58, 103, 104]. It has revealed many important roles played in biological processes such as gene expression regulation [105], development of various human diseases [105-108], drug discovery [109, 110], and biomarker discovery [104]. This section will discuss current RNA sequencing workflows, normalization and downstream analysis.

1.5.3.1 RNA sequencing

To enable NGS analysis, current RNA sequencing protocol involves library generation. Library preparation is the process of converting RNA to cDNA, attaching sequencing-specific adapter sequences and additional motifs such as sequencing binding sites that are complementary to the sequencer. General workflows of library preparation includes: synthesis of cDNA with library indices (barcodes), cDNA cleanup, library amplification, and library clean-up. For non-coding RNA of 22 nucleotides in size, specifically designed 3’ and 5’ adapters are ligated prior to reverse transcription with unique molecular indices (UMI) for efficient ligation. For total RNA sequencing library

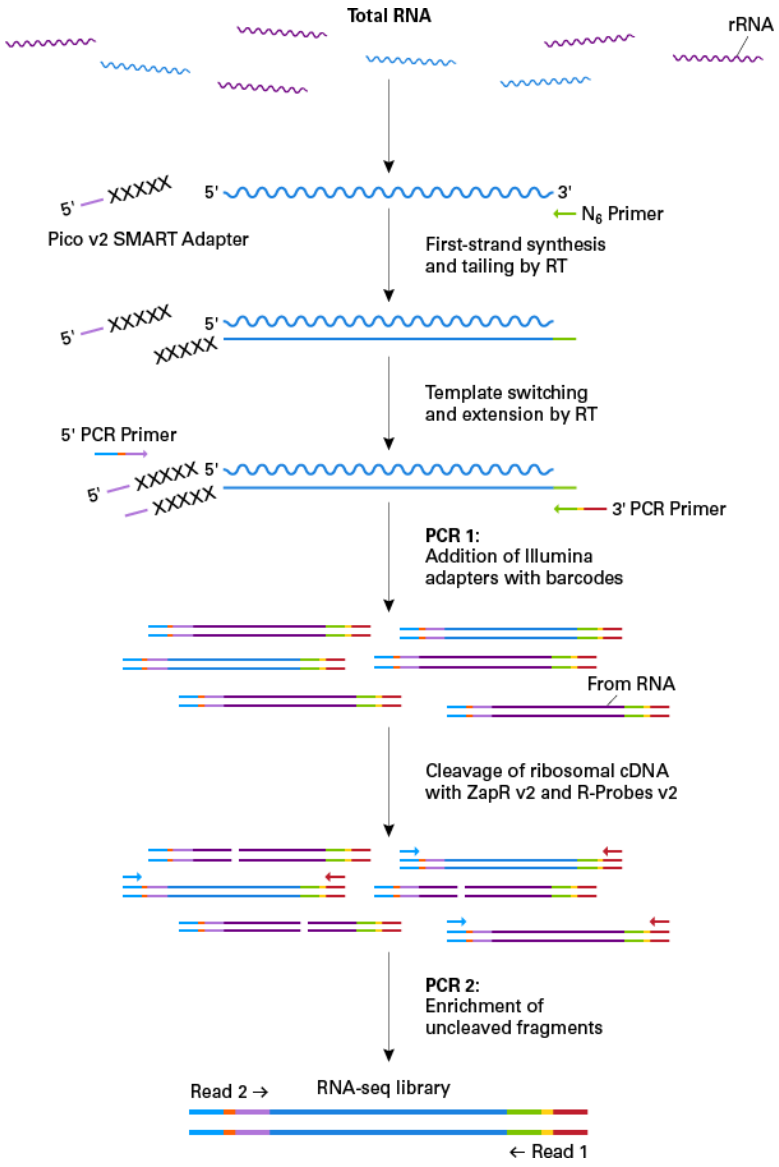


Figure 1.15 | Schematic of SMARTer stranded total RNAseq library preparation

SMARTer stranded total RNAseq involves generation of cDNA from all RNA fragment using random priming (N6 primer). During reverse transcription, enzyme adds few nontemplatated nucleotide shown as Xs to the 3' end of cDNA. The adapter base-pairs with nontemplated nucleotide stretch, creating an extended template to continue replicating to the end of oligonucleotide. The resulting cDNA contains sequencing derived from random primer and adapter. The resulting cDNA goes through 2 PCR cycles which adds full length of Illumina adapters including barcode and enrichment of final library after ribosomal cDNA cleavage. Reprinted with permission from [111]. Copyright Takara Bio Inc. Mountain View, CA, U.S.A.

preparation, important consideration should be taken to cleave highly abundant ribosomal cDNA. This process leaves the library fragments originating from non-rRNA molecules untouched [112]. The final library contains sequences allowing clustering on an Illumina flow cell. Studies have compared the performance of commercially available library kits for samples with low amounts of total RNA [113]. In summary, SMARTer Stranded Total RNA-Seq Kit v2 provided strand specificity while working with minute starting materials, allowing to provide better resolution of transcriptomic profiling [113]. Additionally, SMARTer Stranded total RNA sequencing kits have been successfully utilized for low-input RNA amount from human biofluids (**Figure 1.15**) [103].

1.5.3.2 Workflow for RNA sequencing

Following RNA-sequencing, resulting reads are quality controlled and undergo sequencing alignment to generate a count matrix (**Figure 1.16**). Quality control can be assessed using quality control tools like FastQC. FastQC provides overview of RNA-seq raw reads consisting of sequence quality, GC content, adaptor content, duplicated reads, and overrepresented sequences. The first pipeline starts with a reference alignment against the human reference genome. After the quality of raw reads is assessed, read alignment is performed to determine where the reads originated from in the reference genome. Alignment is then performed by packages like TopHat, STAR (Spliced Transcripts Alignment to Reference), HISAT, and bowtie2 which consists of removing tagging sequence reads, assigning sample reads per unique library barcode, and removing erroneous reads which are either too long or too short [114-118]. Mapping statistics which include statistics of uniquely mapped reads, reads mapped to multiple location, and reads that are unmapped can be obtained through read count distribution. Once the reads are

aligned to the genome, the next step is to generate a count matrix that has been mapped to genome. Ht-seq count and featureCounts are two commonly used counting tools. Once a count matrix is generated, differential expression analysis can be performed using tools like DESeq2 [119], edgeR [120], and limma [121]. All of these packages are available in R, which is language useful for analyzing NGS data.

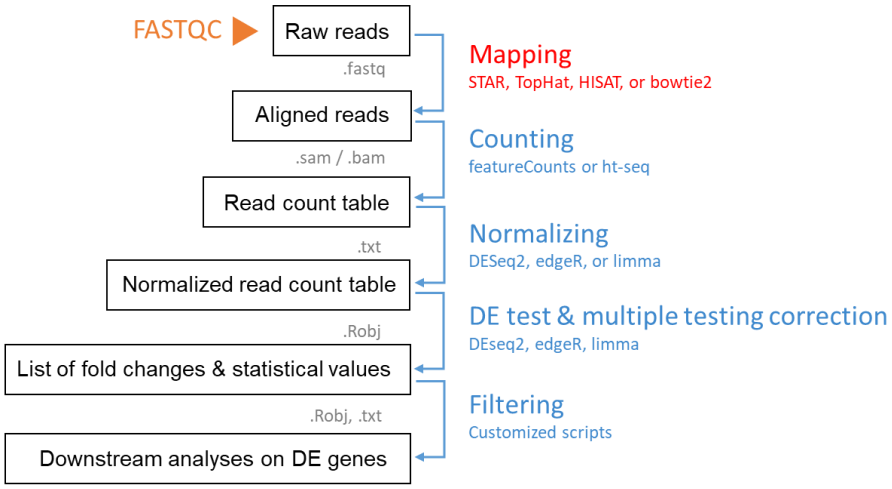


Figure 1.16 | Workflow of RNAseq analysis

The unmapped sequencing file is in FASTQC file format generated from next-generation sequencing technologies. After quality of sequence is assessed with quality matrics, the reads are aligned to human reference genome using alignment tools (STAR, TopHat, HISAT, or bowtie2). Count matrix is generated from aligned reads using counting tools (featureCounts or htseq) followed by differential expression analysis and other downstream analysis.

1.5.3.3 Normalization

Normalization is essential to reduce sources of systematic variation including library size (sequencing depth) or unwanted variation introduced by technical effects. The basic concept of RNA-seq normalization has been accounting for library size, or the total

number of reads from the input library. Methods commonly used for normalization of RNA-seq data are: trimmed mean of m-value (TMM), median, upper quartile (UQ), scale quantile, RPKM, TPM, RPM, and relative log expression (RLE) using R packages from edgeR and Deseq [122, 123]. Trimmed mean of m-value is a normalization method in which scaling factors are calculated based on the weighted mean of log ratios between test and reference [123, 124]. Median normalization works by dividing each count by the median expression of all genes in an observation and multiplying by the median values from all observations. Upper quartile is similar to median normalization, except that the 50% quantile is replaced by a 75% quantile. Scale quantile is user-specified quantile method. RPKM (reads per kilobase per million reads) or TPM (transcript per million) are commonly used in normalization that includes gene length correction [125, 126]. RPM (reads per million mapped reads) are calculated from the number of reads mapped to a gene RPKM x 10^6 divided by the total number of mapped reads [127]. Finally, RLE is calculated using the log ratio of gene counts over the geometric mean across all samples [128]. Although global gene expression analysis provides quantitative information, these methods assume samples have similar total expression [128, 129]. However, potential sources of error can be overlooked when the relative mRNA expression levels significantly differ in biological samples (**Figure 1.17**). Utilizing global scaling factor may lead to either a false increase or regression in the expression level according to the library size. To overcome this issue, synthetic spike-in RNA standards are implemented which allows normalization to correct for unwanted variations [129, 130]. External RNA Controls Consortium (ERCC) spike-in RNA is a premixed set of 92 synthetic transcripts which share common attributes of eukaryotic mRNA, including polyadenylated tails, at differing nucleotide lengths and

concentrations [130]. Utilizing ERCC as an external standard, Loven et al. ensured more accurate detection of differential expression of samples with inherently different biological library sizes [129]. Therefore, normalization should consider variability in measurements by both biological and technical factors in completing RNA-seq studies.

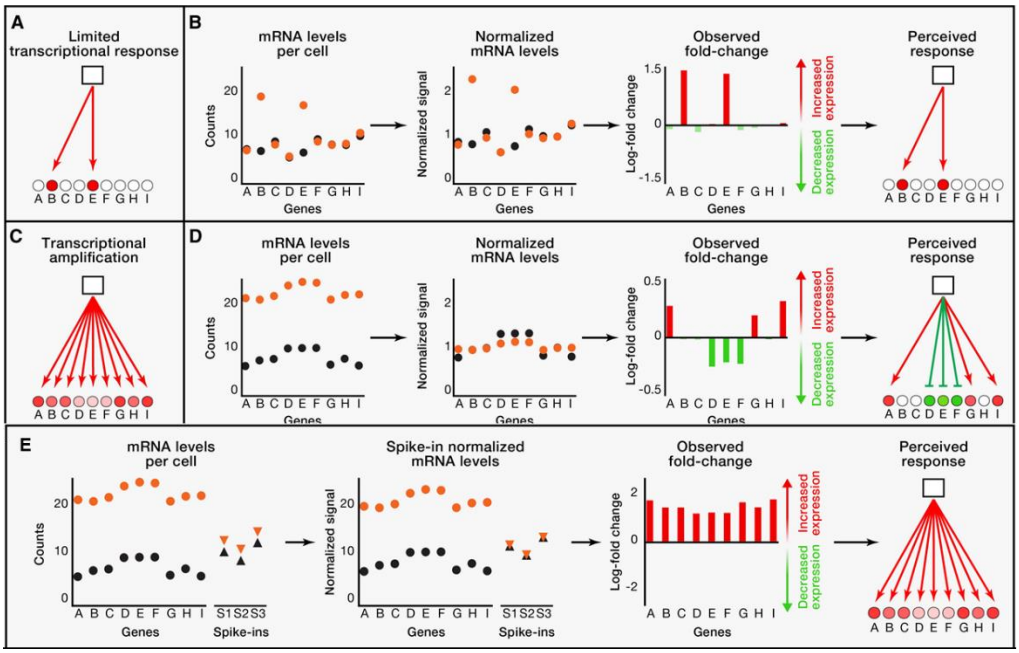


Figure 1.17 | RNAseq normalization and interpretation of expression data

(A) Schematic representation of transcription change on limited sample. (B) Schematic representation of effect of normalization when overall levels of mRNA (in black dots) do not change. The observed fold change reveals the increased expression represented by red bars above the midline and decreased expression represented by green bars. (C) Schematic representation of pattern of change in gene expression when levels of total RNA in two cells are different, where most genes are also expressed in higher level. This will lower overall observed fold change in those samples. (D) Schematic representation of normalization when overall observed mRNA levels per cells are increased. (E) The perceived response using spike-in as standards for normalization reveal overall transcriptional amplification of gene expression. Reprinted with permission from [129]. Copyright 2012 Elsevier.

1.5.3.4 Downstream analysis

Finally, after all other steps to pre-process and normalize sequencing data, expression level differences across sample types can be compared to find differentially expressed (DE) genes of interest. DE genes are commonly defined as any gene which has a collective expression level change lower than a threshold p-value. As with most statistical comparisons, the number of samples and variation within and between replicates all affect DE genes. By organizing DE genes by lowest p-value, most up- or down-regulated genes between samples can be extracted. Genes implicated by DE analysis are often candidates for downstream validation, either by comparison with additional sequencing datasets or even RT-qPCR to more directly quantify how significantly different expression from genes identified through sequencing are across samples. Critically, robust genes which have passed DE analysis are often good candidates for either diagnostic or prognostic applications.

1.6 Gaps in current understanding & layout of the thesis

Collectively, the current literature presents key challenges which limit the promise of liquid biopsy based early cancer detection: i) the lack of standardized blood processing for multi-omics which minimizes ex-vivo processing artefacts via discerning true-in-vivo signatures from ex-vivo artefacts; ii) the daily cycles and consumption of meal influence on cell free mRNA and EV levels are not clear; iii) a comprehensive study of cf-mRNA for cancer detection, pan cancer discernment, and high risk group identification has not been conducted; and iv) the selective packaging of cf-mRNA carriers and normalization pipelines for quantitative data analysis are unknown. My central hypothesis is that

circulating cfRNA/EVs contain cancer associated signatures, which can be detected within patient plasma to stratify cancers and precancerous conditions. The remainder of the thesis is organized as follows. Chapter 2 focuses on preanalytical variability assessment on both EVs and cf-RNA, investigating ex-vivo artefacts from blood processing conditions. Chapter 3 aims to reveal diurnal stability of EV and cf-RNA signatures across blood sampling, establishing baseline levels for within and between individuals. Chapter 4 aims to identify specific gene signatures which can differentiate the presence of cancers and precancerous diseases from healthy patients, allowing us to identify lowly abundant yet robust biomarkers. Finally, chapter 5 aims to identify the carrier of cf-mRNA in patient plasma, and determine if selective RNA packaging can reveal the progression of cancer.

Chapter II: Irreversible alteration of extracellular vesicle and cell-free messenger RNA profiles in human plasma associated with blood processing and storage

2.1 Abstract

The discovery and utility of clinically relevant circulating biomarkers depend on standardized methods that minimize preanalytical errors. Despite growing interest of studying extracellular vesicles (EVs) and cell-free messenger RNA (cf-mRNA) as potential biomarkers, how blood processing impacts both remain to be established. To systematically investigate this, we utilized flow cytometric analysis and examined impact of differential centrifugation and freeze/thaw effect on EV profiles. Utilizing flow cytometry post acquistion analysis software (FCMpass) to calibrate light scattering and fluorescence, we revealed how differential centrifugation and post-freeze/thaw processing removes and retains EV subpopulations. Additionally, cf-mRNA levels measured by RT-qPCR profiles from a panel of housekeeping, platelet, and tissue-specific genes were preferentially affected by differential centrifugation and post-freeze/thaw processing. We found this is predominantly due to freezing plasma containing residual platelets, yielding irreversible *ex vivo* generation of EV subpopulations and cf-mRNA transcripts. Importantly, we found distinct subpopulation of EVs and cf-mRNA in human plasma persisted despite additional processing after freeze-thaw, highlighting importance of minimizing confounding variation attributable to plasma processing and platelet contamination.

2.2 Introduction

Circulating extracellular vesicles (EVs) and cell-free RNA (cfRNA) are promising biomarkers for early cancer detection [131]. EVs are a heterogeneous mixture of vesicles with varying size and composition that are released from cells [132-135]. Since EVs are either derived from the plasma membrane or the involvement of multivesicular endosome fusion with the cell surface, they have cell-specific antigens on their surface that may be antibody labelled for imaging and/or isolation in –omics analyses [7, 136, 137]. There is increasing evidence that EVs may transport a variety of proteins and nucleic acids, including being a potential carrier of cfRNA [138, 139]. Cell-free messenger RNAs (cf-mRNA) specifically are protein coding mRNA molecules in plasma that may serve as biomarkers [140, 141]. Since EVs may transport diverse extracellular RNAs, including cf-mRNA, there is an intense interest in the combination of these analytes for blood-based cancer diagnosis [142, 143].

Previous studies have suggested that *ex vivo* platelet activation and fragmentation affect EV profiles in serum and plasma [132, 144-146, 148-151]. Thus, the International Society of Extracellular Vesicles (ISEV) and International Society on Thrombosis and Haemostasis (ISTH) have recommended general platelet-poor plasma processing conditions for EV analysis [16, 152]. However, how specific preanalytical variables can influence EV subpopulations was not thoroughly characterized. Others have shown that residual platelets also significantly affect plasma microRNA levels solely due to differences in blood processing methods [145, 146]. However, no prior studies have specifically examined changes in cf-mRNA. Since common blood processing conditions for biobanking may not produce platelet-poor plasma [145], as guided by ISEV and ISTH,

additional processing on banked samples after thawing may mitigate the effect of platelet activation on EVs and cf-mRNA analysis. However, which subpopulations of *ex vivo* generated EVs and cf-mRNA subtypes are removable or retained is unknown.

Despite the growing body of literature describing the impact of blood processing on EVs, standardization through light scatter calibration was not widely adopted in these studies to analyze EV subpopulations using flow cytometry. Flow cytometry has been increasingly utilized to characterize heterogeneity of EV surface markers [132, 145, 153-158]. Nonetheless, standardizing nanoscale flow cytometry for sub-micron sized EV detection can be challenging due to varying instrument settings and resolution [155-157]. Recent efforts to ameliorate this have focused on Mie scattering theory modeling [40, 41, 49, 50, 159]. An estimated relative size of an EV population can be derived from a given scatter intensity provided an assumed refractive index and specific optical configuration in a flow cytometer [40, 49, 50]. Although quantifying the exact refractive index of EVs is challenging, previous measurement by either nanoparticle tracking analysis or fluorescence lifetime imaging microscopy suggested a potential range from 1.37-1.45 [44, 45, 160]. Using National Institutes of Standards and Technology (NIST) traceable bead standards with known diameters and refractive indices, scatter-diameter curves can be generated via postacqusition analysis software [40, 41, 49, 50, 159]. Given an effective refractive index of EVs, established scatter-diameter curves yielded reproducible EV measurement between instruments [40, 41, 49, 50, 159].

In this study, we systematically examined the variation of both EV and cf-mRNA subpopulations in human plasma due to blood processing and freeze thaw effect after -80°C storage. EVs were analyzed by flow cytometry with standardized size and fluorescent

calibration, and cf-mRNA levels were measured by multiplex RT-qPCR. We compared plasma derived from single spin (S1: 1,000 × g centrifugation for 10 min) and double spin (S2: 15,000 × g secondary spin for 10 min after the initial single spin S1) analyzed freshly and after freezing. We examined how post freeze/thaw processing removes and retains specific EV subpopulations as well as cf-mRNA originated from platelets, common cell types and tissue specific cells. Our analysis revealed subpopulations of EVs and cf-mRNA were irreversibly altered *ex vivo* in association with blood processing and freeze/thaw effects after storage.

2.3 Materials and Methods

Blood sample collection and processing

All experimental protocols were reviewed and approved by the Oregon Health & Science University Institutional Review Board. All methods were carried out in accordance with relevant guidelines and regulations. Blood samples from healthy individuals were obtained from the Cancer Early Detection Advanced Research center (CEDAR) at Oregon Health and Science University. All samples were collected under institutional review board (IRB) approved protocols with informed consent from all participants for research use. Whole blood was collected from healthy individuals in 10 ml in K2EDTA tubes (BD Vacutainer, Becton Dickinson, cat. 36643) via antecubital vein puncture using a 21G butterfly needle (BD Vacutainer, Becton Dickinson, cat. 367281). Tubes were transported vertically at room temperature before processing. Within 1 hour of blood withdrawal, 10 ml of whole blood was centrifuged at 1,000 × g for 10 minutes at 23°C with the highest acceleration and deceleration setting at ‘9’ using Eppendorf 5810-R centrifuge with S-4-

104 Rotor. Plasma was collected until 10 mm above the buffy coat and was labelled as S1. To obtain double spun plasma, S1 plasma was centrifuged in Eppendorf 5424R centrifuge at 15,000 x g for 10 minutes at 23°C. The resulting supernatant of platelet-depleted plasma was collected and labelled as S2. S1 and S2 plasma samples did not undergo a freeze/thaw cycle. Plasma samples that were frozen at −80 °C and thawed at room temperature were labelled as S1FR and S2FR respectively. For post-thaw processing, S1FR was centrifuged in Eppendorf 5424R centrifuge at 15,000 x g for 10 minutes at 23°C. The resulting supernatant was carefully transferred and designated as S1FRS2.

Platelet counting

The platelet count was measured by an improved Neubauer haemocytometer (VWR Scientific Products, Piscataway, NJ) by two independent, experienced researchers. The total number of platelets were counted from central 1 x 1 mm area consisting of 25 groups of 16 squares separated by closely ruled triple lines, equivalent to a volume of 0.1 µl.

Flow cytometry set-up for light scatter and fluorescence calibration

Beckton-Dickinson FACSAria Fusion equipped with 488 nm (60 mW) and 640 nm (100 mW) lasers was used. For optimal configuration of submicron size detection, 0.1 µm size filter was applied to the sheath fluidic system to reduce sheath fluid noise. The sample flow rate was set at 1, which was measured by mass discharge [159] and determined to be 45 µl/minute. Timed collections were recorded for 60 seconds. Data collection was set using the SSC trigger threshold value of 200 using scatter wavelength at 488 nm. In order to calibrate light scattering, 152, 203, 303, 401, 510, and 600 nm polystyrene NIST-

traceable beads (ThermoFisher Scientific) were serially diluted in 0.1 μm filtered D-PBS without calcium and magnesium. Minimum of 5,000 events were recorded for 60 seconds. Particle diameter and scatter relationship was established utilizing FCMpass software (v3.09, http://nanopass.ccr.cancer.gov) [40, 49, 50]. Median SSC-H intensity in arbitrary units was converted to standardized unit in EV diameter. To approximate EV diameter size, the average of effective refractive index (RI) data based upon published measurements were used. Detailed instructions for light scattering calibration based on a core-shell structure to model EVs were followed (Shell RI = 1.4800, Core RI: 1.3800, and shell thickness: 5 nm) [40]. For fluorescence calibration, Quantum Alexa Fluor 647 Molecule Equivalent Soluble Fluorochrome (MESF) (Bangs Laboratories, cat. 647) or Quantum Alexa Fluor 488 MESF (Bangs Laboratories, cat. 488) were used. Data collection was set using the FSC trigger threshold value of 5,000 and analyzed using FSC-A vs SSC-A in arbitrary units. Utilizing FCMpass software, the fluorescent intensity in arbitrary units were converted to MESF standardized units. All measurements were analyzed using FlowJo software.

Fluorescent antibody labeling of differentially processed plasma for flow cytometry

To fluorescently label EV surface proteins, 5 μl of plasma was incubated with 5μl of antibody mix prepared after established dilution series. CD9 Alexa Fluor 647 (R&D system, clone: #209306, cat. FAB1880R-100μg) was diluted to a final concentration of 0.001 mg/ml for staining. CD63 Alexa Fluor 488 (Thermofisher scientific, clone: MEM-259, cat. MA5-18149, concentration 0.26 mg/ml) was diluted to a final concentration of 0.0013 mg/ml for staining. For isotype controls, mouse IgG2B Alexa Fluor 647 conjugated

isotype control (R&D system, cat. IC0041R) and mouse IgG1 Alexa Fluor 488 conjugated isotype control (Thermofisher Scientific, cat. MA518167) were used at the same concentration as matched stained controls and were recorded at the same dilution as stained and unstained samples. Incubation was done for 3 hours at room temperature in the dark. The stained EV samples were further diluted 200-fold with 0.1 µm filtered D-PBS without calcium and magnesium Thermo Fisher Scientific, cat. 14190250) prior to acquiring the data using an abort rate of < 5% and keeping the threshold rate below 20,000 events per second. To account for the electronic abort rate due to nanoparticle coincidence (also known as "swarming"), stained samples were serially diluted and validated via consistent median fluorescent intensity across plasma dilutions. A buffer-only control of 0.1 µm-filtered DPBS without calcium and magnesium was recorded at the same flow cytometer acquisition settings as all other samples, including triggering threshold, voltages, and flow rate. The buffer-only control had a count of <1,000 events per second.

RT-qPCR profiling of cell free mRNA

For characterizing the effect of freeze thaw on cell free mRNA expressions, RNA was extracted using plasma processed with S1, S2, S1FR, S2FR, and S1FRS2 conditions. Cell free mRNA was isolated by using plasma/serum circulating and exosomal RNA purification Kit (Norgen Biotek) followed by 10X Baseline-ZERO DNase treatment (Epicentre). DNase treated RNA samples were purified and further concentrated using RNA clean and concentrator (Zymo Research). The purified RNA samples were assayed by RT-qPCR using custom selected 16 primers targeting MTND2, PPBP, B2M, PF4, ACTB, CORO1C, GSE1, GAPDH, SMC4, HBG1, NUSAP1, MIKI67, FGB, APOE, FGG,

and ALB. Template RNA was mixed with Superscript III One-step RT-PCR system with Platinum Taq DNA polymerase (Invitrogen) to generate cDNA according to the protocol. PCR amplification products were treated with Exonuclease I (New England Biolabs) to digest single stranded primers at 37°C for 30 min followed by inactivation of enzymes at 80°C for 15 min. For RT-qPCR, cDNA from preamplification was diluted 1:80 and set-up in 96-well plates with SsoFast EvaGreen supermix with low ROX (BioRad) with above primers at 10μM. QuantStudio 7 Flex (Applied Biosystems) was used to run RT-qPCR assay according to manufacturer's recommended cycling conditions.

Statistical analysis

To determine the impact of overall preanalytical factors, statistical analysis was performed on $CD9^+$ or $CD63^+$ EVs on specific gated populations across differential centrifugation and freeze/thaw processing. The significance of individual preanalytical factor comparisons were determined using Tukey's multiple comparison test. p values <0.05 were considered statistically significant (*p < 0.05, **p < 0.01, ***p < 0.001, and ****p < 0.0001). Analyses were conducted using R package.

2.4 Results

Light scattering and fluorescence calibration

We calibrated the flow cytometer Beckton-Dickinson FACSAria Fusion using National Institutes of Standards and Technology (NIST) traceable size standard beads (152, 203, 303, 401, 510, and 600 nm) and Quantum Molecules of Equivalent Soluble Fluorophore (MESF) to establish standardized units for light scatter and fluorescence

respectively [40, 49, 50]. For light scatter, each bead sample was analyzed at the same acquisition setting until > 5,000 bead events were recorded. The histogram of each sized bead population revealed distinct side scattering in arbitrary units, where a progressive increase in SSC-H with increasing NIST bead diameter (152 nm – 600 nm) was observed (**Figure 2.1A**). The median light scatter statistic of each bead size gated population was inputted into flow cytometry post acquisition analysis software (FCMpass) to calibrate light scattering [40, 49, 50]. The collection half-angle of our system, which is important to quantify the amount of light reaching a detector in absolute units, was determined to be 45.3° using FCMpass software. Utilizing the side scattering collection angle, recorded side scattering value in arbitrary units was standardized to predicted scattering cross-section using Mie theory [40, 49, 50]. The linear regression between our observed light scattering power in arbitrary units and predicted scattering cross-section resulted in R-squared value = 0.9991 (**Figure 2.1B**). The acquired scattering intensity of standard beads (red dots) was plotted on modelled data (black line) for polystyrene beads, which revealed the model fitted actual data accurately (**Figure 2.1C**). After scatter-diameter relationship for EVs is extrapolated using FCMpass software, the measured scatter signal for polystyrene beads corresponding to vesicle diameter is revealed (**Figure 2.1C**). Approximate diameters of EVs was calculated using the average of effective EV refractive index (Shell RI = 1.4800, Core RI: 1.4000, and shell thickness: 5 nm) [40, 49, 50].

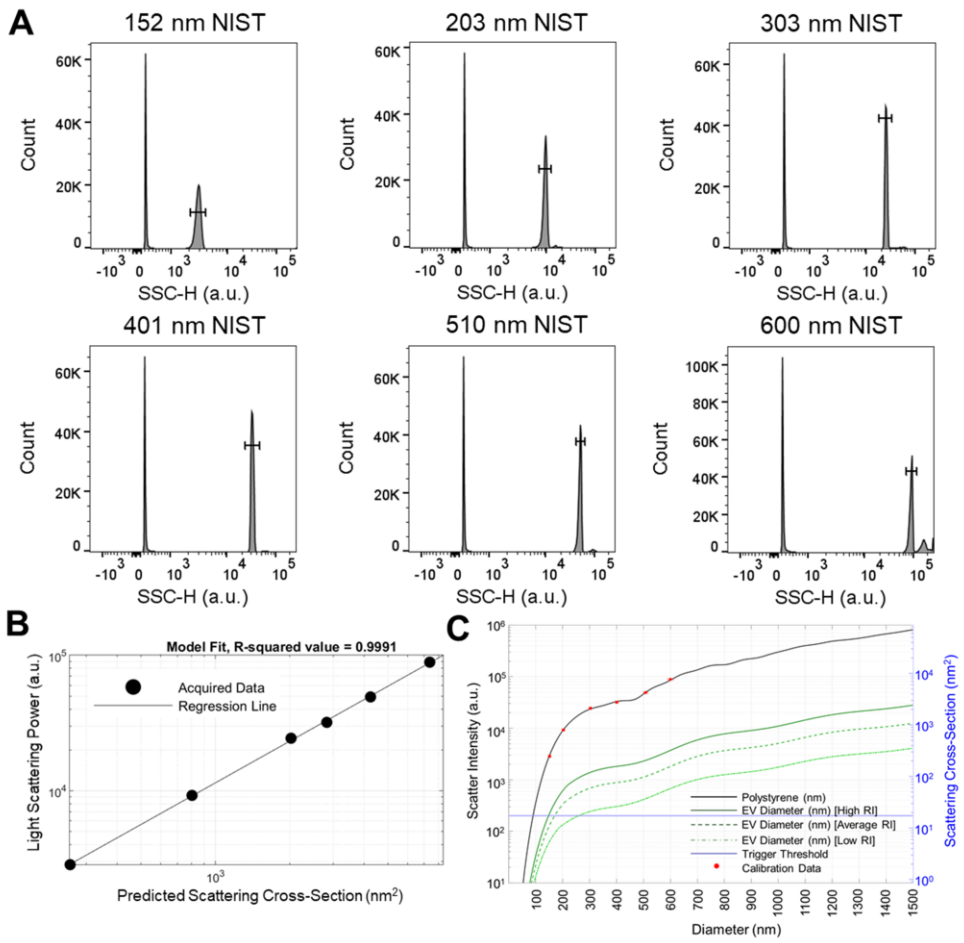


Figure 2.1 | Light scattering calibration

(A) Histogram of NIST-traceable polystyrene beads (152, 203, 303, 401, 510, and 600 nm) are shown using side scattering (SSC-H) on a bioexponential scale. Each bead size relative to SSC-H is identified to obtain median side scattering in arbitrary units (a.u.) for light scattering calibration. (B) Regression plot of acquired light scattering power in arbitrary units compared to the predicted scattering cross-section in nm^2 is calculated using FCMpass software. (C) Scatter-diameter curve showing light scatter intensity relationships with EV diameter established in FCMpass software. The acquired NIST-traceable polystyrene bead scattering intensity are overlaid with the predicted scattering data for NIST-traceable polystyrene beads with refractive index of 1.5900. The scatter-diameter relationship given high, average, and low effective EV refractive indices are shown, which can be used to estimate EV diameter from corresponding scattering intensity in arbitrary units.

Next, we performed fluorescence calibration using forward scatter as the trigger threshold to gate micron-sized MESF beads. MESF beads for each fluorophore was analyzed until > 5,000 bead events were recorded. While FSC-A vs. SSC-A revealed a single microsphere population, four fluorescent microspheres were observed with varying fluorescent intensity in arbitrary units (**Figure 2.2A, 2.2B**). The median fluorescence statistic of each bead fluorescence gated population was inputted into flow cytometry post acquisition analysis software (FCMpass) to calibrate fluorescence (**Figure 2.2C**). The relationship between MESF bead reference values and acquired fluorescence in arbitrary units was established to calibrate fluorescence (**Figure 2.2D**).

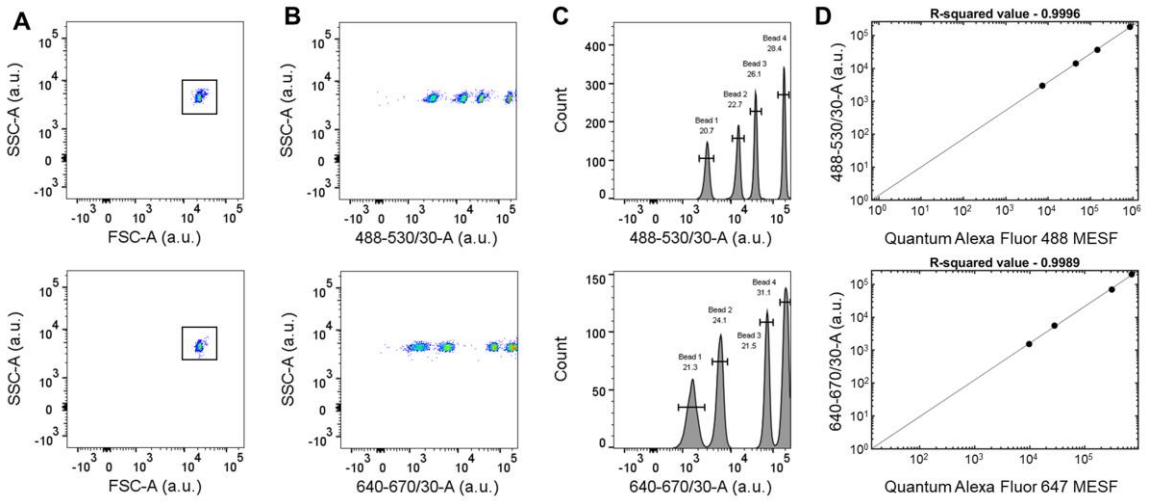


Figure 2.2 | Fluorescence calibration

(A) Representative flow cytometry dot plots of Quantum Alexa Fluor 488 MESF (top) and Quantum Alexa Fluor 647 MESF beads (bottom) gated using SSC-A and FSC-A in arbitrary units (a.u.). (B) The gated beads are shown in each fluorescence channel (488-530/30-A and 640-670/30-A respectively) against SSC-A in arbitrary units. (C) Histogram of Quantum Alexa Fluor 488 MESF and 647 MESF beads are shown using fluorescent intensity in arbitrary units. (D) Regression of acquired fluorescence intensity in arbitrary units to MESF bead reference values for each population established in FCMpass software.

Flow cytometry reveals distinct vesicle populations differentially affected by blood processing condition

After establishing light scatter and fluorescence calibration, we investigated the impact of differential centrifugation on plasma EVs using the flow cytometer. Plasma was differentially processed using single centrifugation at 1,000 × g for 10 min (S1) and double centrifugation (S2: 15,000 × g secondary spin for 10 min after the initial single spin S1) (**Figure 2.3A**). Complete counts of residual platelets in plasma were measured using a hemocytometer. Single spun plasma S1 contained an average platelet concentration of 313 ± 74 thousand/µl while secondary spin resulted in the removal of more than 99.99% of residual platelets in S2 (**Figure 2.3B**). For EV analysis, plasma was stained with anti-CD9 and anti-CD63 fluorescent antibodies and measured by flow cytometry. The fluorescently positive gated data revealed that there are distinct populations in EV diameter distribution ranging between 150 nm and 3,000 nm (**Figure 2.3C**). It is noted that the subset of EVs within 150 – 1,000 nm range at around 500 nm is an artifact of Mie scattering calibration from our calculated flow cytometer collection angle and geometry. Specifically, this corresponds to a plateau in the scatter-diameter curve from ~400-480 nm using predicted EV light scattering from the estimated average EV refractive index employed in our model (**Figure 2.1C**). Welsh et al. reported a similar observation, suggesting that a plateau from the scatter-diameter curve resulted in an artifact between 400 – 480 nm accordingly [40]. Therefore, we gated EVs into two populations: 150 – 1,000 nm which may be comprised of small and medium EVs, and 1,000 – 3,000 nm comprised of large EVs and platelets [159]. Flow cytometer assay controls included unstained samples, isotype controls, serial dilution of stained plasma, and antibody with buffer alone (**Supplementary Figure S2.1**).

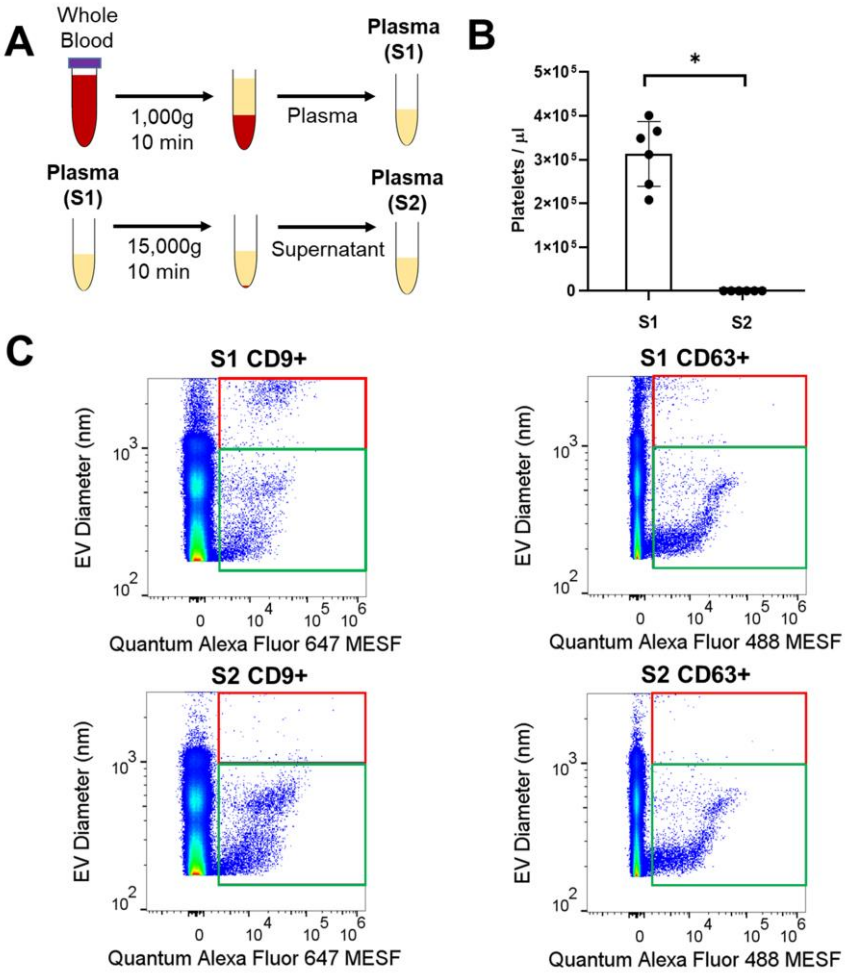


Figure 2.3 | Effect of differential centrifugation on EVs using flow cytometry

(A) Schematic diagram of differentially processed plasma using single spin (S1: 1,000 × g centrifugation) and double spin (S2: 15,000 × g secondary spin after the initial single spin S1). (B) Platelet concentration in differentially processed plasma from three healthy individuals (n=3) was measured in independent technical replicates using a haemocytometer. Values are means ± standard deviations for the indicated blood processing conditions. (C) Representative flow cytometry dot plot of EV diameter (nm) versus fluorescent intensity in Quantum Alexa Fluor MESF units for S1 and S2. Quantum Alexa Fluor 647 MESF was used for Alexa Fluor 647 conjugated CD9 stained plasma, and Quantum Alexa Fluor 488 MESF was used for Alexa Fluor 488 conjugated CD63 stained plasma. Events were gated into two subpopulations: 150 to 1,000 nm (green box) and from 1,000 to 3,000 nm (red box).

Serial dilution of stained plasma showed the linear detection of EVs while the median fluorescence intensity remained constant, suggesting that EVs were detected and counted as single particles via flow cytometry. Plasma condition at S2 resulted in a clear reduction in 1,000 – 3,000 nm population compared to S1 while the 150-1,000 nm EVs remained similar for plasma from both processing conditions (**Figure 2.3C**).

Freezing of platelet containing single spun plasma generates ex vivo EVs

Previous studies have suggested that ex vivo platelet activation and fragmentation generate EVs [15-17, 132]. We noted that single spun plasma S1 contained a high level of residual platelets (**Figure 2.3B**). Therefore, we examined freeze/thaw effect on plasma EV profiles using anti-CD9 and anti-CD63 fluorescent antibodies. We compared EVs measured freshly (S1 and S2) with EVs after a freeze/thaw cycle (S1FR and S2FR) (**Figure 2.4A, Supplementary Figure S2.2**). We observed remarkably increased CD9+ EVs for both 1,000 – 3,000 nm and 150 – 1,000 nm populations after single freeze/thaw cycle of S1 plasma (S1FR vs. S1, $P < 0.001$). However, we observed no significant changes in CD9+ EVs occurred after single freeze/thaw cycle of S2 plasma samples for either size (S2FR vs. S2, ns) (**Figure 2.4B**). Similarly, the 1,000 – 3,000 nm CD63+ EVs were significantly increased in single spun plasma after freeze/thaw (S1FR vs. S1, $P<0.01$) while remaining the same for double spun plasma (S2FR vs. S2, ns) (**Figure 2.4B**). In contrast, the 150 – 1,000 nm CD63+ EVs were statistically unchanged with respect to either spin freeze/thaw cycle (**Figure 2.4B**). To confirm the nature of EVs which are sensitive to detergent lysis, we applied detergent to disrupt EVs found in S1FR and S2FR. We found disappearance of both CD9+ and CD63+ stained EVs with detergent treatment, validating the detected

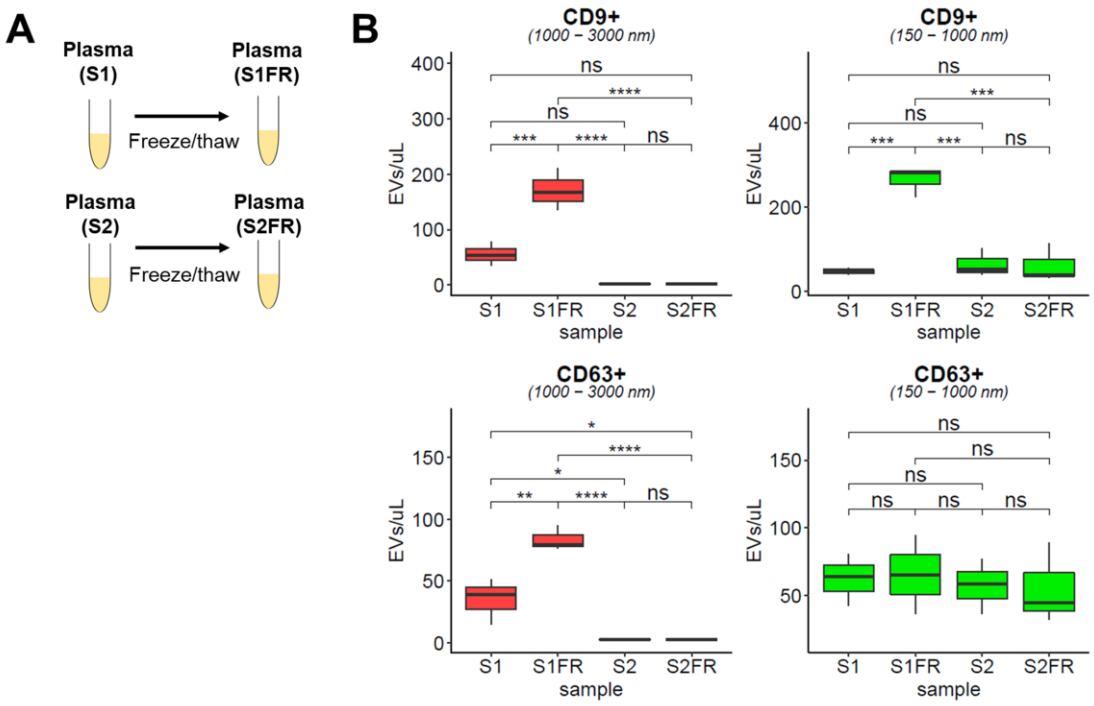


Figure 2.4 | Effect of freeze thaw cycle on EVs using flow cytometry

(A) Schematic diagram of differentially processed plasma (S1, S2) and respective freeze thaw processes (S1FR, S2FR). (B) Box plot of CD9+ and CD63+ of gated events from 1,000 – 3,000 nm (red) and 150 – 1,000 nm (green) for S1, S1FR, S2, and S2FR. CD9+ and CD63+ events were converted to concentrations using calibrated flow rate in a given acquisition time. EV concentration defined as the number of EVs per µl was determined by number of EVs detected in a given sample volume. Statistical significance obtained from three healthy volunteers for each freeze thaw processing using Tukey's multiple comparisons (ns = not significant, P>0.05; *P <0.05, ***P<0.001, ****P<0.0001).

difference is not due to false-positive events derived from antibody aggregates (**Supplementary Figure S2.2**). In summary, our data indicated that freezing single spun plasma which contains residual platelets generated ex vivo EVs in a marker dependent manner, whereas no significant change was observed for residual platelet depleted plasma in the second spin prior to freezing.

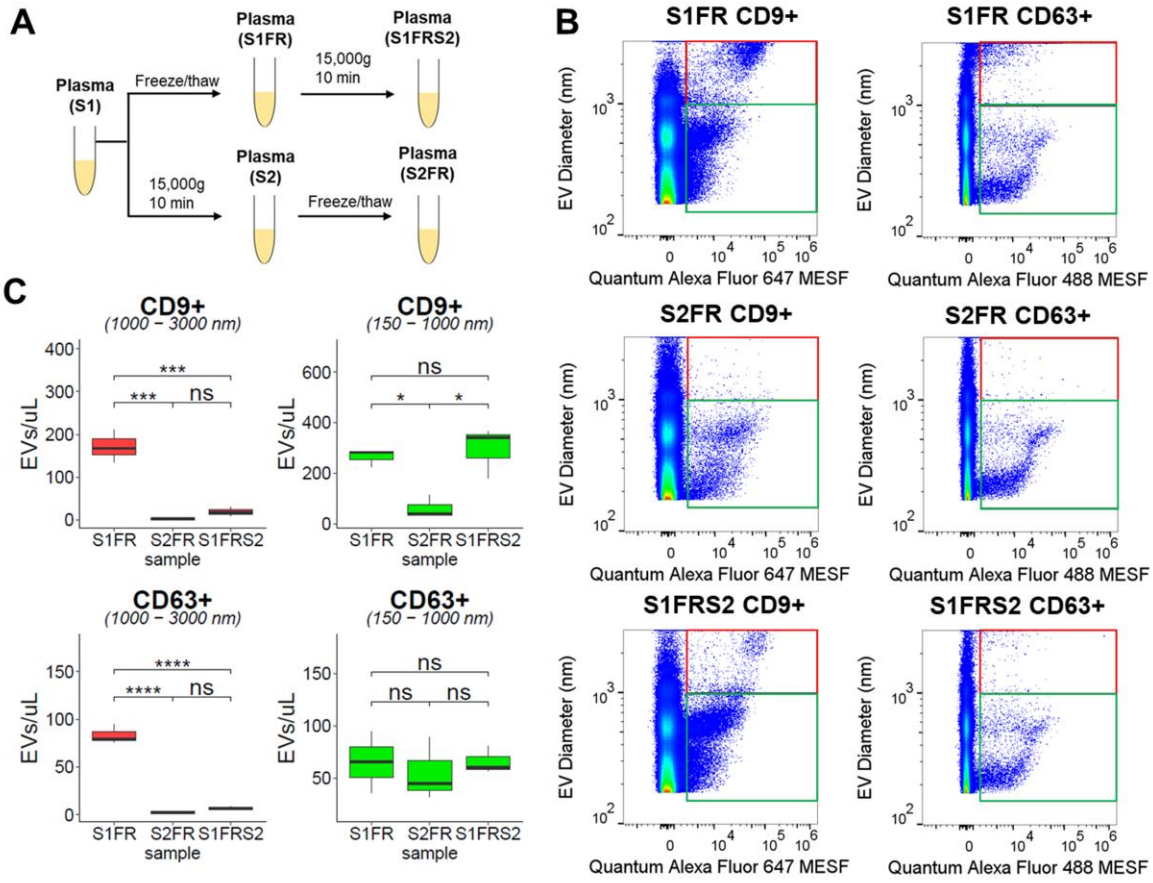


Figure 2.5 | Effect of post-thaw processing on EVs using flow cytometry

(A) Schematic diagram of differentially processed plasma (S1, S2), respective freeze thaw samples (S1FR, S2FR), and secondary spin after post freeze/thaw plasma S1FR (S1FRS2). (B) Representative flow cytometry dot plot of EV diameter (nm) versus fluorescent intensity in Quantum Alexa Fluor MESF units for CD9+ EVs and CD63+ EVs in S1FR, S2FR, and S1FRS2 conditions. Quantum Alexa Fluor 647 MESF is used for Alexa Fluor 647 conjugated CD9 stained plasma, and Quantum Alexa Fluor 488 MESF is used for Alexa Fluor 488 conjugated CD63 stained plasma. Events were gated from 150 to 1,000 nm (green box) and from 1,000 to 3,000 nm (red box). (C) Box plot of CD9+ and CD63+ of gated events from 1,000 – 3,000 nm (red) and 150 – 1,000 nm (green) for S1FR, S2FR, and S1FRS2. Statistical significance were obtained from three healthy volunteers for each freeze thaw processing condition using Tukey's multiple comparisons (ns = not significant, P>0.05; *P <0.05, ***P<0.001, ****P<0.0001).

Ex vivo generated EVs are irreversible even after post-thaw processing

Next, to test if a post-thaw processing effectively removes ex vivo generated EVs, we performed centrifugation at 15,000 g for 10 min on S1FR plasma samples (S1FRS2) (**Figure 2.5A**). We found S1FRS2 significantly depleted CD9+ and CD63+ 1,000 – 3,000 nm populations associated with S1FR (**Figure 2.5B, 2.5C**). Meanwhile, the levels of small and medium CD9+ EVs associated with S1FR remained significantly higher in post-thaw processing plasma S1FRS2 compared to S2FR (S1FRS2 vs S2FR, $P<0.05$ for EV diameter 150 – 1,000 nm) (**Figure 2.5B, 2.5C**). In contrast, we observed 150 – 1,000 nm CD63+ EVs remained statistically unchanged (S1FRS2 vs S2FR, ns) (**Figure 2.5B, 2.5C**). Collectively, our results revealed freezing residual platelets in S1 significantly generated small and medium CD9+ EVs ex vivo, which post-thaw processing could not remove. Meanwhile, we observed CD63+ EVs were retained regardless of spinning and post-thaw processing conditions.

Distinct subsets of cf-mRNA influenced by differential centrifugation and post-thaw processing

Since platelets and EVs contain mRNA, we sought to determine if blood centrifugation and post-thaw processing affected cf-mRNA levels. We analyzed cf-mRNA profiles in single and second spin plasma (S1 and S2) freshly, after freezing at -80°C (S1FR and S2FR), and for samples subjected to a second spin following S1FR processing (S1FRS2). We selected a panel of housekeeping, platelet and tissue-specific genes for multiplex RT-qPCR measurements (**Figure 2.6A, 2.6B**). Hierarchical clustering analysis of relative gene expression between post-thaw processed samples revealed three distinct

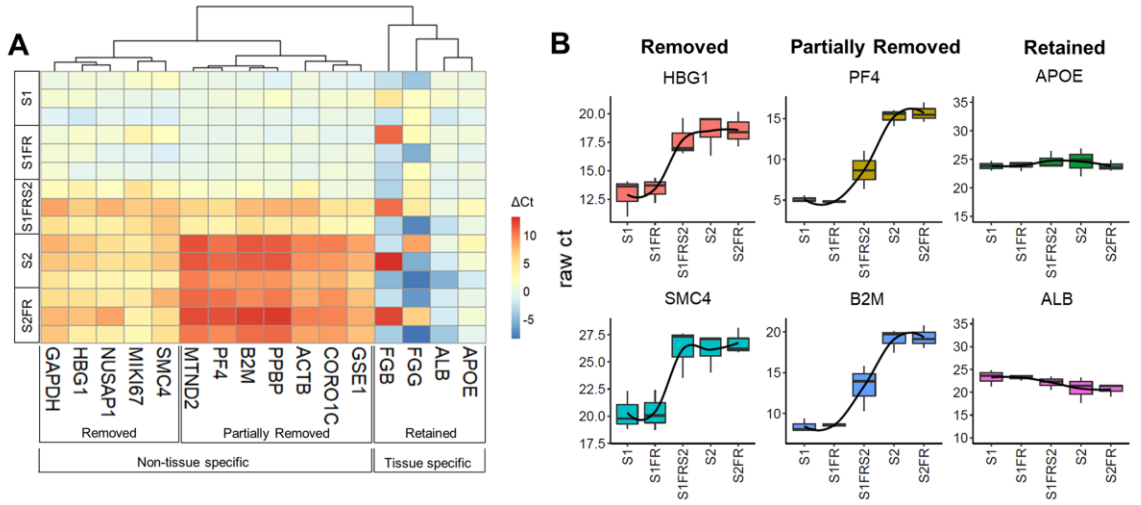


Figure 2.6 | Effect of freeze thaw and post-thaw processing on cf-mRNAs using qRT-PCR

(A) Hierarchical clustering analysis of relative levels (in ΔCt) of 16 custom selected genes using RT-qPCR. Ct difference (ΔCt) between S1 and individual processing conditions are indicated from lowest (blue) to highest (red). Non-tissue specific genes that are fully or partially removed, and tissue-specific genes which are retained in S1FRS2 with respect to S1 are shown. (B) Box plot of the median expression levels (in Cts) for representative non-tissue specific genes which are fully removed (i.e. HBG1 and SMC4) or partially removed (i.e. PF4, and B2M), and tissue-specific genes (i.e. APOE, ALB) which are retained in S1FRS2 with respect to S1 are shown. Higher raw Ct indicates lower levels of cfRNA transcripts.

clusters (**Figure 2.6A**). Overall, these clusters were either dependent (non-tissue specific) or independent (tissue specific) of post-thaw processing conditions, wherein non-tissue specific genes segregated into two clusters. The first cluster included genes (HBG1 and SMC4 for example), which could be removed by post thaw processing and therefore were likely related to large EVs or platelets (**Figure 2.6A, 2.6B**). The second cluster, including platelet genes and house-keeping genes such as PF4 and B2M, was partially removed by

post thaw processing and therefore was likely associated with ex vivo generated small and medium EVs which remained after post-thaw processing (**Figure 2.6A, 2.6B**). Importantly, our results revealed that tissue specific gene signatures (such as genes expressed in liver tissue; including APOE and ALB) were retained regardless of spinning and post-thaw processing conditions (**Figure 2.6A, 2.6B**), suggesting they are present in non-platelet small or medium EVs. The relationship of cf-mRNA transcripts with EV subpopulations requires further investigation and is the subject of future studies. Overall, as genes from different biological roles are uniquely affected by preanalytical differences, the selection of novel cfRNA biomarkers should consider the effects of preanalytical variability.

2.5 Discussion

Circulating EVs and cf-RNA are promising biomarkers for disease diagnosis and prognosis [161-164]. However, significant variability in standardizing blood processing across published methods has led to a lack of reproducibility between studies [132, 145, 146, 165]. In this study, we utilized multiparametric flow cytometry and cf-mRNA profiling to characterize preanalytical influences on EV and cf-mRNA subpopulations in plasma. We observed two distinct subpopulations by flow cytometry which are differentially impacted by centrifugation and post-thaw processing. Interestingly, we observed small and medium $CD9^+$ EVs irreversibly generated via freezing single spun plasma while $CD63^+$ EVs remained similar. Importantly, these *ex vivo* generated EVs could not be removed by additional centrifugation after freeze/thaw, and thereby can significantly affect downstream analyses. As a first in cf-mRNA studies, we also found groups of genes significantly, partially, or unaffected by post-thaw processing in plasma.

Since different types of EV purification methods (ultracentrifugation, density gradients, size-filtration, etc.) affect the yield and purity of EVs [22, 166], we chose to fluorescently label EVs directly in plasma using EV tetraspanin-specific antibodies with proper assay controls according to recent MIFlowCyt-EV reporting framework [167]. Previous study highlighted effects of centrifugation on pre-isolated EVs from platelets and erythrocytes by examining recovery of EVs through differential centrifugation [18]. However, their freeze-thaw cycle was performed on purified EVs, leading to no significant change across different temperatures of single freeze-thaw cycle. By examining EVs directly in plasma, we observed irreversible *ex vivo* generation of EVs and cf-mRNA subpopulations altered by differential centrifugation and post-thaw processing.

Although previous studies highlighted the preanalytical influences on microparticle generation associated with platelet activations [17, 18, 168], the effect of blood processing on EV subpopulations using flow cytometry with light scattering and fluorescence standardized calibration is lacking. The enumeration of microparticles in previous studies mostly utilized flow cytometry that was validated to discriminate between 0.5 μm and 0.9 μm Megamix beads [15, 18]. Since considerable efforts have been directed to establish a standardized methodology for EV measurements by flow cytometry [40, 41, 49, 50, 159], we applied this standardized approach to investigate enumeration of EVs influenced by preanalytical factors. Utilizing the FCMpass software developed by Welsh et al. [40, 49, 50], we observed differential centrifugation results in distinct EV subpopulations within the diameter range between 150 nm and 3,000 nm. In addition, our freeze thaw analysis further revealed *ex vivo* generated $CD9^+$ EVs, adding to the body of literature on platelet-associated blood processing artefacts [132, 145]. We addressed the importance of

preanalytical influences on EVs using a standardized approach, which will improve the reproducibility with respect to effective EV diameter and given fluorochrome molecule standards across literatures.

Comprehensive assessment of light scattering sensitivity on multiple different flow cytometers was performed by Van del pol et al [159]. For small particle detection, only a few flow cytometers detected more than three different sized reference beads using both side scatter (SSC) and forward scatter (FSC). Similarly, our instrument could not detect more than three sized reference beads by FSC. Instead, we utilized FCMpass software to calibrate SSC using the effective refractive index of EVs (Shell RI = 1.4800, Core RI: 1.3800, and shell thickness: 5 nm) [40]. Since the true refractive index of different EV subpopulations is currently unknown, the average EV refractive indices based on core-shell theory has been implemented [40, 49]. Although EVs in the ~ 1,000 nm diameter range may overlap with small platelets [169], precise refractive indices which considers platelet granule content and shapes is currently unknown. Specific studies, which definitively parse EVs from small platelets and understanding refractive indices of EV subpopulations, are needed to better define EV physical characteristics and compositions.

How blood processing influences circulating microRNA has been previously shown [145, 146], and yet the impact on cf-mRNA is poorly understood. Cheng et al. provided preanalytical influences on miRNA expression due to differing residual platelet amount [146]. Conversely, our study investigated the impact of blood processing conditions through differential centrifugation, respective freezing condition, and post-thaw processing on cf-mRNA. We revealed cf-mRNA groups whose extent of preanalytical variability differed based on the degree of residual platelets in plasma. In particular, non-

tissue specific genes were further classified as either partially or fully removed by freeze-thaw post processing. Intriguingly, tissue-specific cf-mRNA were less prone to blood processing conditions, revealing them as potentially more robust biomarkers, or differentially associated with smaller vesicle subpopulations retained through centrifugation.

2.6 Conclusions

In conclusion, our study provides an assessment of the preanalytical effect of differential centrifugation and freeze/thaw cycles on plasma EVs and cf-mRNA. Employing multiparametric flow cytometry, our work provides insights into how preanalytical factors influence EV subpopulations and *ex vivo* release of EVs in association with residual platelets. Notably, these artifacts appear to be irreversible for $CD9^+$ small and medium EVs and mRNA transcripts of genes present in platelets. Our results indicate distinct subpopulations of EVs and cf-mRNA are not removable by additional spinning after freeze/thaw. Therefore, consideration should be taken when analyzing EVs and cf-mRNA from banked plasma and designing robust EV and cf-mRNA based liquid biopsy tests.

Acknowledgements

The authors would like to acknowledge Joshua A. Welsh from National Cancer Institute at NIH for helpful discussion and implementation of FCMpass software. We are also grateful to Nick Wang of OHSU for facilitating access to Quantstudio 7 Flex Real-Time PCR systems. We are also grateful to Jeong Yoon Lim for biostatistics advice and the Cancer Early Detection Advanced Research (CEDAR) center of the OHSU Knight Cancer Institute for assistance with obtaining blood samples. The authors also thank Pamela S Canaday, Brianna Garcia, and Dorian Latocha for flow cytometry training at Oregon Health and Science University. We acknowledge funding support from Cancer Early Detection Advanced Research (CEDAR) center at Oregon Health & Science University's Knight Cancer Institute, Cancer Research UK/OHSU Project Award (C63763/A27122) and by grants from the National Institutes of Health (R01HL101972, R21HD16-037, R01GM116184 and R01HL047014). The authors declare no competing financial interests.

Author Contributions

TN conceived and supervised the project. HK, FC, SY, and TN developed the initial workflow and established the blood processing optimization on EVs, platelets and cf-mRNA. TKM and MM developed the flow cytometry assay in our group. HK, MR, FC, TKM and TN developed the analysis strategy. HK, MR, SY, RA, FC, and TN acquired and performed data analysis. HK and TN wrote the manuscript. TKM, MR, SY, FC, RA, AN, and OM edited the manuscript. All authors have reviewed and approved the manuscript.

Chapter two, in part, is a reprint (with co-author permission) of the material as it appears in the submitted manuscript: "Irreversible alteration of extracellular vesicle and cell-free messenger RNA profiles in human plasma associated with blood processing and storage", Hyun Ji Kim, Matthew Rames, Samuel Tassi Yunga, Randall Armstrong, Mayu Morita, Owen McCarty, Fehmi Civitci, Terry Morgan, Thuy Ngo, Submitted to *Scientific Reports* (2021). The author of this dissertation is the primary author of this manuscript.

Chapter III: Diurnal stability of cell-free DNA, cell-free RNA, and extracellular vesicles in human plasma samples

3.1 Abstract

Many emerging technologies are reliant on circulating cell-free DNA (cfDNA), cell-free RNA (cfRNA), or extracellular vesicles (EVs) for applications in the clinic. However, the impact of diurnal cycles or daily meals on circulating analytes are poorly understood and may be confounding factors when developing diagnostic platforms. To begin addressing this knowledge gap, we obtained plasma from four healthy donors serially sampled five times during 12 hours in a single day. For all samples, we measured concentrations of cfDNA and cfRNA using both bulk measurements and gene-specific digital droplet PCR. In addition, we measured the abundance of plasma EVs immunostained with canonical EV and platelet markers using flow cytometry. We found no significant variation attributed to blood draw number for the cfDNA, cfRNA, or EV measurements throughout the day. This indicated that natural diurnal cycles and meal consumption do not appear to significantly affect abundance of total cfDNA, total cfRNA, our two selected cfRNA transcripts, or common EV markers in plasma. Conversely, we observed significant variation between individual donors for cfDNA, one of the cfRNA transcripts, and two of the EV markers. The results of this work suggest that it will be important to consider patient-specific baselines when designing reliable circulating cfDNA, cfRNA, or EV clinical assays.

3.2 Introduction

Liquid-biopsy based diagnostic platforms are a highly desired and increasingly accepted in clinical settings [170]. Although many diseases could potentially benefit from liquid-biopsy technology, the need for non-invasive platforms is especially apparent in the

field of cancer screening and diagnostics because of potential risks involved with invasive needle biopsy procedures (e.g. [171, 172]) and concerns of radiation exposure during imaging tests [173, 174]. In addition, there is substantial interest in accurate screening methods that can be performed at frequent intervals to stratify patient cancer risk and detect potentially lethal cancers at early, treatable stages [175]. Plasma or serum-based platforms are of particular interest because the circulatory system interacts with the entire body and therefore provides a means to sample all organs. The sensitivity of current methods to reliably characterize nucleotide sequences and subcellular particles at single-event resolution has generated substantial interest in utilizing circulating cell-free DNA (cfDNA, e.g. [176-179]), cell-free RNA (cfRNA, e.g. [142, 180-183]) and extracellular vesicles (EVs, e.g. [138, 184]) as clinically-relevant biomarkers. Human plasma contains both vesicular and extravesicular RNA and DNA; these different components may have distinctive contents with potential clinical relevance [185]. While promising, translation of cfDNA, cfRNA, and EVs to the clinic has been slow in part because the natural temporal and interpersonal variation of the circulating analytes remains poorly understood. It is well-known that mammalian blood cell/tissue gene expression and physiology changes drastically during the daily diurnal cycle or following meals [186-189]. Recent evidence also suggests that some human bodily fluid-derived micro-RNAs (miRNAs) may follow a daily cycle of fluctuation [190, 191]. Therefore, establishing the extent to which circulating cfDNA, cfRNAs, and EVs are affected by normal physiology is critical for the analytes to have successful clinical implementation.

To address the potential influence of daily cycles on circulating analytes, we characterized total cfDNA, cfRNA, and EVs in plasma obtained from four healthy

volunteers sampled multiple times across two days. We measured bulk cfDNA and cfRNA concentration using fluorometric or automated electrophoresis methods, as well as sequence-specific cfDNA and cfRNA copy number concentration using digital droplet PCR (ddPCR). The droplet-counting approach of ddPCR allows for absolute quantitation of nucleic acid templates and more sensitive resolution of fold-changes compared to conventional quantitative PCR [192-194]. Using this ddPCR approach, we targeted two single-copy genomic DNA regions as proxies for cfDNA abundance: one locus containing the gene *telomerase reverse transcriptase* (*TERT*) and one locus containing the gene *N-acetylglucosamine kinase* (*NAGK*); for cfRNA we targeted two commonly used genes used for mRNA normalization: *β-actin* (*ACTB*) and *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*). Using high-resolution flow cytometry following International Society for Extracellular Vesicles (ISEV) guidelines [195], we quantified EVs immunostained for canonical exosomal markers CD9, CD81, and CD63 and the platelet marker CD41 [55, 196]. Our results suggest that while cfDNA, cfRNA, and EVs are overall stably expressed diurnally, several of the analytes demonstrate significant interpersonal or daily variation. Deeper characterization of these sources of variation will likely be required before the circulating analytes gain greater acceptance as clinically practical liquid-biopsy platforms.

3.3 Materials and Methods

Participant plasma sample collection

Identifier	Age (years)	Sex
HD1	42.5	F
HD2	50	M
HD3	60.1	F
HD4	73.7	F

Table 3.1 | Age and sex of the four healthy donors (HDs) volunteered for this study.

All experimental protocols were reviewed and approved by the Oregon Health & Science University Institutional Review Board (protocol #8316). All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all volunteers, and volunteers were compensated for participating. Healthy donors (HDs) were consented for multiple blood draws over a 12-hour period at the Oregon Health & Science University Oregon Clinical and Translational Research Institute (OCTRI) inpatient research clinic. HD age and sex distributions are given in **Table 3.1**. Each HD had two days, separated by one week, to provide five blood draws each day (**Figure 3.1**). The HDs were advised to not engage in rigorous exercise 24 hours prior to the blood draw dates. HDs were given access to recliner chairs and had the ability to walk around freely between blood draws. All individuals had an IV inserted for the multiple blood draws, but at times when the IV failed, ad hoc venipuncture was performed about 50% of the time for each participant. Blood was drawn from individuals every 2 hr 45 min beginning at 8:30 am. Meals ordered from the hospital menu were consumed by individuals between draw 1-2, draw 2-3, and draw 4-5. To replicate a typical patient arriving in a clinical setting, diets

were not restricted. Approximately 20 ml of blood was drawn into 10 ml EDTA tubes (Cat# 366643, BD Vacutainer) per time point. Blood was processed within 15 min of the draw and plasma was obtained as follows: EDTA tubes were centrifuged at 1,000 x *g* for 10 min at room temperature, plasma was extracted down to ~500 µl from the buffy coat interface, plasma supernatant was centrifuged for a second time at 2,500 x *g* for 10 min at room temperature, and the resulting plasma supernatant was extracted down to ~200 µl from the debris pellet interface. The final supernatant was distributed into 1 ml aliquots and immediately frozen at -80°C until analysis.

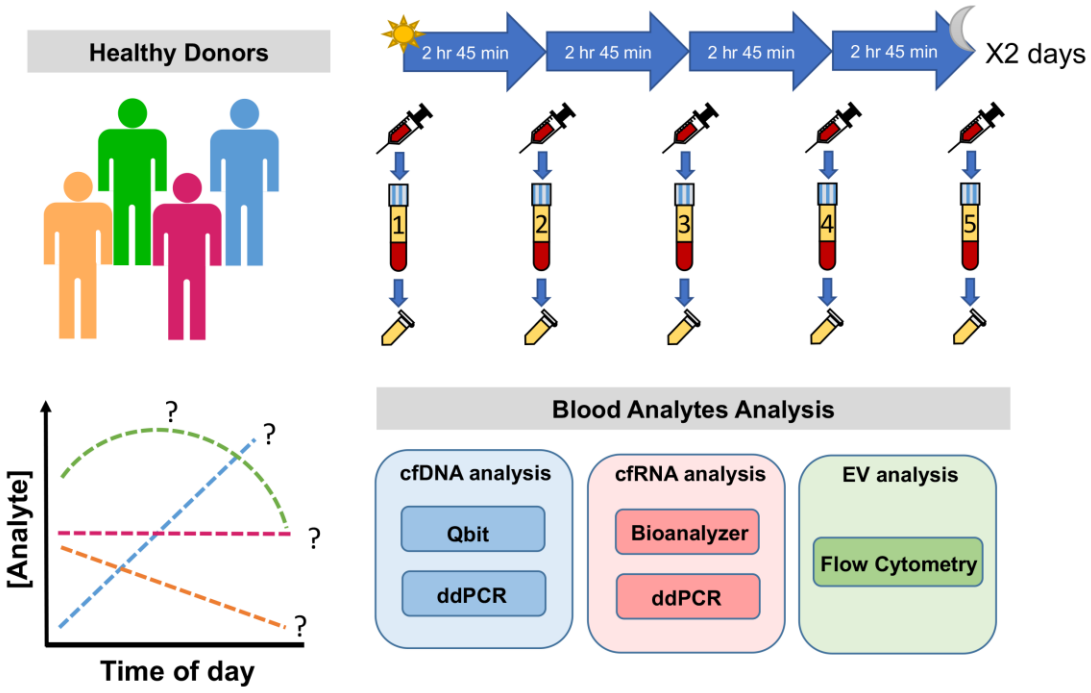


Figure 3.1 | Schematic of the HD sampling procedure used to obtain plasma for analysis.

Blood from four HDs were sampled five times per day over two days. The blood was processed into plasma and analyzed for changes in cfDNA, cfRNA, or EVs.

Nucleic acid extractions

For each HD and time point, cfDNA and cfRNA was extracted separately (**Figure 3.1**). CfDNA was extracted from 1 ml plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen #55114) according to manufacturer instructions and eluted into 20 µl buffer EB (10 mM Tris-Cl, pH 8.5). CfRNA was extracted from 1 ml plasma using the Plasma/Serum Circulating and Exosomal RNA Purification Kit (Norgen #42800) and eluted into 100 µl nuclease-free deionized water. To remove genomic DNA contamination, the cfRNA samples were treated with 2 MBU of Baseline-ZERO DNase in 1X Baseline-ZERO DNAse buffer (Lucigen #DB0715K) for 20 min at 37°C, purified using the RNA Clean & Concentrator-5 kit (Zymo #R1013), and eluted in 14 µl nuclease-free water. For no plasma controls, 1 ml nuclease-free deionized water was used in place of plasma. These purified cfDNA and cfRNA samples were used for all subsequent nucleic acid analyses.

Bulk quantitation of cfDNA and cfRNA

Purified cfDNA concentration was first determined using the Qubit dsDNA HS Assay Kit (quantification range 0.2 – 100 ng DNA, Thermo Fisher Scientific #Q32854) and Qubit 3 Fluorometer (Thermo Fisher Scientific). Purified cfRNA concentration was quantified using the Agilent RNA 6000 Pico Kit (quantification range 50 – 5000 pg/µL RNA in water, Agilent #5067-1513) and Agilent 2100 Bioanalyzer instrument (Agilent) within a window of 50-500 bp.

First-strand cDNA synthesis of cfRNA

For first-strand cDNA synthesis of cfRNA, 10 µl reverse transcription reactions were prepared using 3 µl total cfRNA in 1X SuperScript IV VILO Mastermix (Thermo Fisher Scientific #11756050). cDNA synthesis reactions were incubated at 25°C for 10 min, followed by 50°C for 10 min, and were terminated with incubation at 85°C for 5 min. The cDNA reactions were used as direct template for cDNA copy number quantification by ddPCR.

Primers for ddPCR analysis

Primers and probes for cfDNA ddPCR analysis to target single-copy number genes *TERT* (F primer 5'-3': CCTCACATAAATGCTACCAAACGA; R primer 5'-3': TTCCAAGAAGGAGGCCATAGTC; Probe 5'-3': AAGAAATGAACAGACCCATC CCCCAGG; fluorescent probe: HEX; quencher: ZEN/IBFQ) or *NAGK* (F primer 5'-3': TGGGCAGACACATCGTAGCA; R primer 5'-3': CACCTTCACTCCCACCTCAAC; Probe 5'-3': TGTTGCCCGAGATTGACCCGGT; fluorescent probe: FAM; quencher: ZEN/IBFQ) and were purchased from IDT (Integrated DNA Technologies, Coralville, IA, USA). Primer and probe sequences for cfDNA were chosen using sequences reported by Devonshire et al. [197] for these two genomic loci. To quantify cDNA copy number, gene expression ddPCR assays for *ACTB* (assay dHsaCPE5190199; fluorescent probe: FAM) and *GAPDH* (assay dHsaCPE5031597; fluorescent probe: HEX) were purchased from Bio-Rad and contained both probes and primers premixed at 10X concentration.

ddPCR of cfDNA and cDNA samples

To measure cfDNA copy number by ddPCR, 22 µl ddPCR reaction mixtures were prepared using 2.2 µl purified cfDNA in 1X ddPCR Supermix for Probes (No dUTP) (Bio-Rad #1863024) and with final primer/probe concentrations of 0.9 µM/0.25 µM or 0.2 µM/0.1 µM for *TERT* and *NAGK*, respectively. Each cfDNA ddPCR reaction was multiplexed for both *TERT* and *NAGK*. To measure cDNA copy number by ddPCR, 22 µl ddPCR reaction mixtures were prepared using 1.5 µl undiluted cDNA template in 1X ddPCR Supermix for Probes (No dUTP) (Bio-Rad #1863024), 1X *ACTB* gene expression ddPCR assay mix, and 1X *GAPDH* gene expression ddPCR assay mix. For no template controls (NTCs), 1 µl nuclease-free water was used instead of cfDNA or cDNA template. Reactions were performed in semi-skirted 96-well plates (Eppendorpf #951020362). Plates for droplet generation were heat-sealed with pierceable foil (Bio-Rad #1814040), vortexed briefly, then spun down using a tabletop plate spinner. Droplet generation was performed using a QX200 AutoDG Droplet Digital PCR System (Bio-Rad) with Automated Droplet Generation Oil for Probes (Bio-Rad #1864110) and DG32 Automated Droplet Generator Cartridges (Bio-Rad #1864108). Droplets were deposited into a clean 96-well plate held in a pre-chilled cold block to prevent evaporation. Plates were then heat-sealed with pierceable foil and PCR was performed in a Bio-Rad C1000 thermocycler using the following temperature conditions: 95°C for 10 min, 40 cycles of 94°C for 30 sec followed by 60°C for 1 min, 98°C for 10 min, and then cooling to 4°C until droplets were read. Droplets were counted using the QX200 Droplet Reader (Bio-Rad) using manufacturer's instructions. Positive and negative droplets were subsequently analyzed using QuantaSoft Analysis Pro (v1.0.596, Bio-Rad, Hercules, California, USA).

Fluorescent labeling of EVs with antibodies

To fluorescently label EV surface proteins, 5 µl of plasma was incubated with 5 µl of antibody mix prepared by 1:200 dilution of anti-CD9 (human) Alexa Fluor 647 conjugated antibody (R&D Systems #FAB1880R-100UG), 1:50 dilution of anti-CD81 (human) PE conjugated antibody (clone M38, Thermo Fisher Scientific #A15781), 1:50 dilution of anti-CD63 (human) Alexa Fluor 488 conjugated antibody (clone MEM-259, Thermo Fisher Scientific #MA5-18149), and 1:20 dilution of anti-CD41 (human) Brilliant Violet 421 conjugated antibody (clone HIP8, BioLegend #303730) for 3 hrs at room temperature in the dark.

Flow cytometry analysis of EVs

EV flow cytometry analysis was performed using the BD FACSAria Fusion (BD Biosciences). The threshold value was set at SSC of 200 at flow rate of 1, stopping time at 60 sec, and events to record at 1,000,000 events. To account for electronic abort rate, plasma samples incubated with antibody mix were diluted to retain threshold rates below 20,000 events per second. The following detector settings were used throughout the experiment: 350V for FSC, 365V for SSC, 700V for laser emitting at 488 nm, 680V for laser emitting at 640 nm, 535V for laser emitting at 405nm, and 655V for laser emitting at 561 nm. To run size calibration beads, megamix-plus SSC and FSC (BioCytex) containing submicron sized fluorescent beads (100, 160, 200, 240, 300, and 500 nm) were reproducibly measured prior to every experiment with laser emitting at 488 nm, where the side scattering voltage was adjusted to match the side scattering intensity of 10^4 for 200 nm beads as a size reference.

Statistics and plots

Graphs of cfDNA, cfRNA, and EVs measured over time were prepared using R (v.3.6.1) and Rstudio (v.1.2.5019). Permutation tests for each analyte were performed in Rstudio using the "coin" R package and default parameters [198, 199]. For significant permutation tests, post-hoc pairwise permutation tests were performed using the R package "rcompanion" (v. 2.3.2, http://rcompanion.org) with the "fdr" p-value adjustment method. To test for a significant difference between draws performed on day 1 versus day 2, a permutation test of symmetry was performed on the draws paired by day. To test for a significant difference between draws or between individuals, the values between day 1 and day 2 for each draw were averaged and one-way permutation tests of independence was performed using either draws or individuals as factors. Summary statistics, correlation plots, Spearman nonparametric correlation coefficient, and two-tailed correlation P-value analysis were prepared using GraphPad Prism (v8.3.0, GraphPad Software, San Diego, CA, USA). For all statistical tests, P-values were determined to be significant at a threshold of $P \leq 0.05$.

3.4 Results and Discussion

Over the past decade, it has become increasingly clear that detecting unpredictable disease via blood biopsy, especially at the earliest stages, will require intricate understanding of naturally occurring circulating biomarker variation. Workflow standardization for liquid-biopsy based analytes is the first step towards identifying and minimizing sources of non-biologically relevant variation [200-202]. However, confounding factors related to meals, time of day, and the intrinsic interpersonal variation

could critically affect analyte abundance, normalization, and multi-omic integration. To begin addressing these concerns, here we provide the first descriptions of diurnal cfDNA, cfRNA, and EV measurements derived from the same cohort.

Differences in plasma cfDNA abundance across donors is attributed to interpersonal variation

First, we characterized total donor plasma cfDNA abundance using Qubit as well as total genome copy numbers via locus-specific ddPCR assays. Following cfDNA extractions of plasma from the four HDs, we observed overall averages of 3.05 ng cfDNA / ml plasma when measured by Qubit (SD = 1.2 ng cfDNA / ml plasma, **Figures 3.2A**, **3.2B**, **Supplementary Table S3.1**). When cfDNA was measured by ddPCR, we observed 758.6 copies/ml plasma (SD = 286.2) and 723.5 copies/ml plasma (SD = 299.3) using TERT and NAGK probes, respectively (**Figure 3.2C-3.2F**). The ddPCR workflow for cfDNA yielded an average of 14,658 accepted droplets (range = 10,426 – 18,130; SD = 1,366) per ddPCR reaction well (**Supplementary Figure S3.1**). No plasma controls and NTCs for TERT/NAGK ranged from 0 – 9.3 copies/ml (**Supplementary Figure S3.2**). We observed a strong correlation between the two cfDNA extraction replicates (Spearman rs = 0.73, $P < 0.0001$, **Supplementary Figure S3.3A**) and between the two analysis methods (Spearman rs = 0.97, $P < 0.0001$, **Supplementary Figure S3.3B**). We found no significant difference in cfDNA abundance between the five draws when measured by either Qubit or ddPCR (**Table 3.2**, **Supplementary Tables S3.1-3.3**). This finding contrasts with a recent report by Madsen et al. [203] which found a decrease in plasma cfDNA concentration at their final draw when five draws were performed three hours apart.

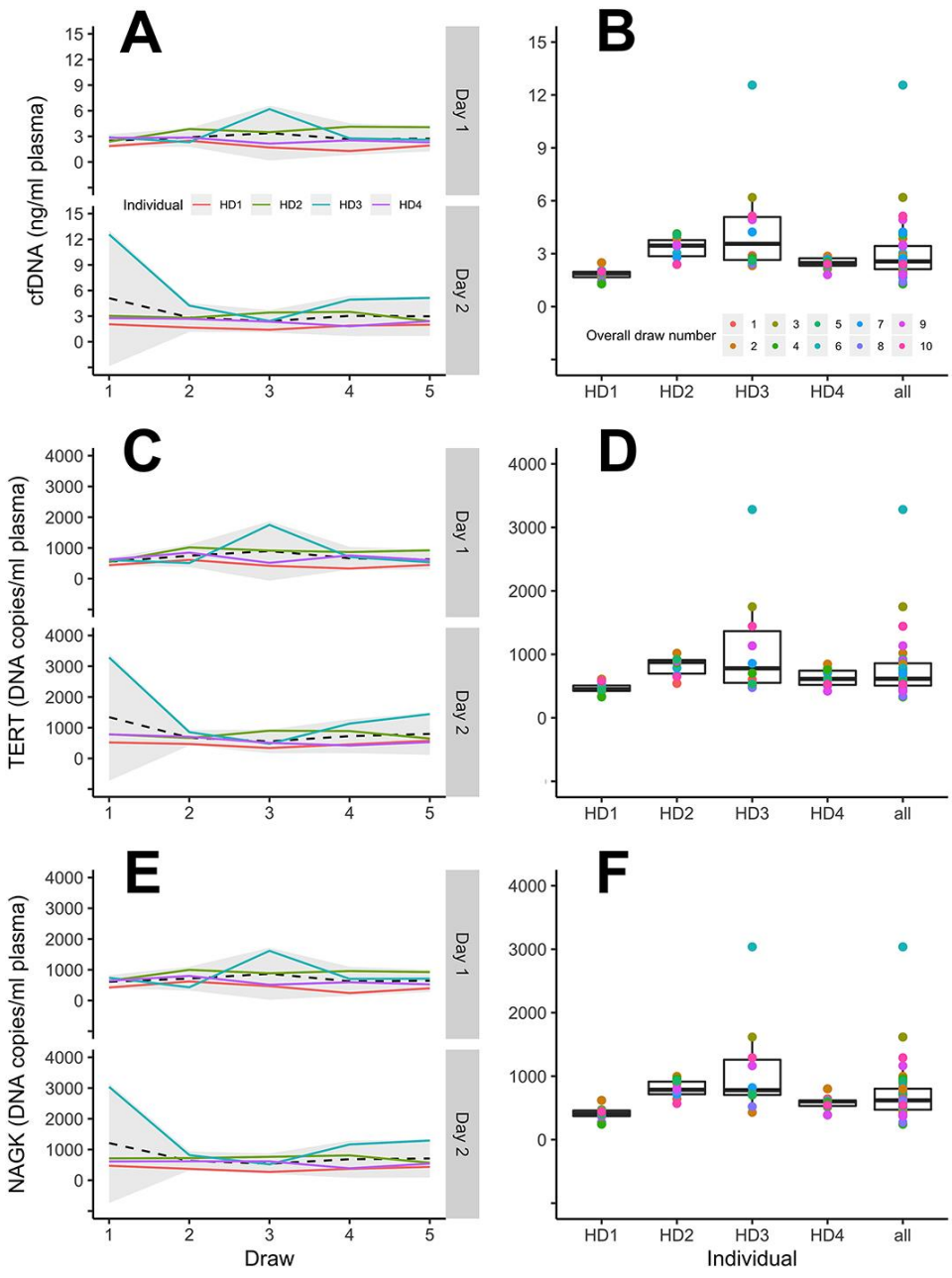


Figure 3.2 | Abundance of plasma-derived cfDNA across the five sampled time points. CfDNA abundance was measured by Qubit (A,B) and by ddPCR with TERT (C,D) or NAGK probes (E,F). For left panels, dashed lines represent the average of the four individuals and the shaded area corresponds to 95% confidence intervals. For right panels, the individual data points overlaying boxplots are color coded sequentially starting from the first blood draw of day 1.

However, our second centrifugation step was performed at 2,500 *x g*, rather than 13,000 *x g* as done by Madsen et al. [203], and therefore the composition of cell-free plasma may not be directly comparable. We observed no significant difference between the two draw days (**Table 3.2**), although we did observe a significant source of variation attributed to the individuals ($P < 0.05$, **Table 3.2**, **Supplementary Table S3.1**).

Analyte	Method	Day[a]	Draw number[b]	Individual[b]
cfDNA	Qubit	0.30 (ns)	0.38 (ns)	0.013 (*)
	TERT	0.61 (ns)	0.38 (ns)	0.021 (*)
	NAGK	0.35 (ns)	0.38 (ns)	0.020 (*)
cfRNA	Bioanalyzer	0.75 (ns)	0.64 (ns)	0.82 (ns)
	ACTB ddPCR	0.96 (ns)	0.80 (ns)	0.20 (ns)
	GAPDH ddPCR	0.29 (ns)	0.53 (ns)	0.016 (*)
EV	CD81+ counts	0.20 (ns)	0.39 (ns)	0.0014 (**)
	CD63+ counts	0.15 (ns)	0.92 (ns)	6.6e-05 (****)
	CD41+ counts	0.90 (ns)	0.96 (ns)	0.059 (ns)
	CD9+ counts	0.50 (ns)	0.99 (ns)	0.076 (ns)

Table 3.2 | Summary of the permutation tests comparing the two draw days, the five draws across the day, or individuals to determine significant sources of variation.

[a] Permutation test of symmetry. P-values are followed by P-value summaries in parenthesis. ns, not significant.

[b] One-way permutation test of independence. P-values are followed by P-value summaries in parenthesis. *, $P < 0.05$; **, $P < 0.01$; ****, $P < 0.0001$; ns, not significant.

For example, the cfDNA level of individual HD1 was consistently lower than the overall average, respectively (**Figure 3.2A-3.2B**, **Supplementary Tables S3.1-S3.3**). Post-hoc pairwise permutation tests also identified individuals with significantly different cfDNA levels (**Supplementary Tables S3.1-S3.3**). Plasma cfDNA was previously described by Zhong et al. [204] to fluctuate 1.9 – 67.9 fold in healthy, nonpregnant

individuals when sampled across 12-hour or longer time points. Similar to Zhong et al., we observed inconsistent cfDNA fluctuation across time in the HDs and attribute primary source of plasma cfDNA abundance differences to be from interpersonal variation.

GAPDH counts, but not ACTB counts or total cfRNA, varied significantly by donor

Next, we measured total plasma cfRNA abundance by Bioanalyzer and mRNA-specific abundance characterization using ddPCR. Across the four HDs, we observed an overall average of 3.05 ng/ml plasma cfRNA when measured by Bioanalyzer (SD = 1.2 ng/ml plasma, **Figures 3.3A**, **3.3B**, **Supplementary Table S3.4**). We did not find significant cfRNA variation by day, draw, or individual when total cfRNA abundance was measured using this method (**Table 3.2**). When cfRNA was measured by ddPCR, we observed 25,022 copies/ml plasma (SD = 5,932) and 5,983 copies/ml plasma (SD = 1,703) using *ACTB* and *GAPDH* probes, respectively (**Figures 3.3C-3.3F**). The ddPCR for cfRNA yielded an average of 16,216 droplets per well (range = 11,912 – 19,618; SD = 1,554) for *ACTB* and *GAPDH* cDNA templates (**Supplementary Figure S3.4**). No plasma controls and NTCs for ACTB/GAPDH ranged from 0 – 4.3 copies/ml (**Supplementary Figure S3.5**). Similar to our bulk measurement of cfRNA by Bioanalyzer, we did not observe significant variation due to day of draw or draw number for either *ACTB* or *GAPDH* (**Figures 3.3C-3.3F**; **Table 3.2**, **Supplementary Tables S3.5-S3.6**). However, we observed significant variation attributed to the individuals for *GAPDH* counts ($P < 0.05$, **Table 3.2**). Post-hoc pairwise permutation tests for *GAPDH* counts did not reveal significant differences between individuals (**Supplementary Tables S3.6**).

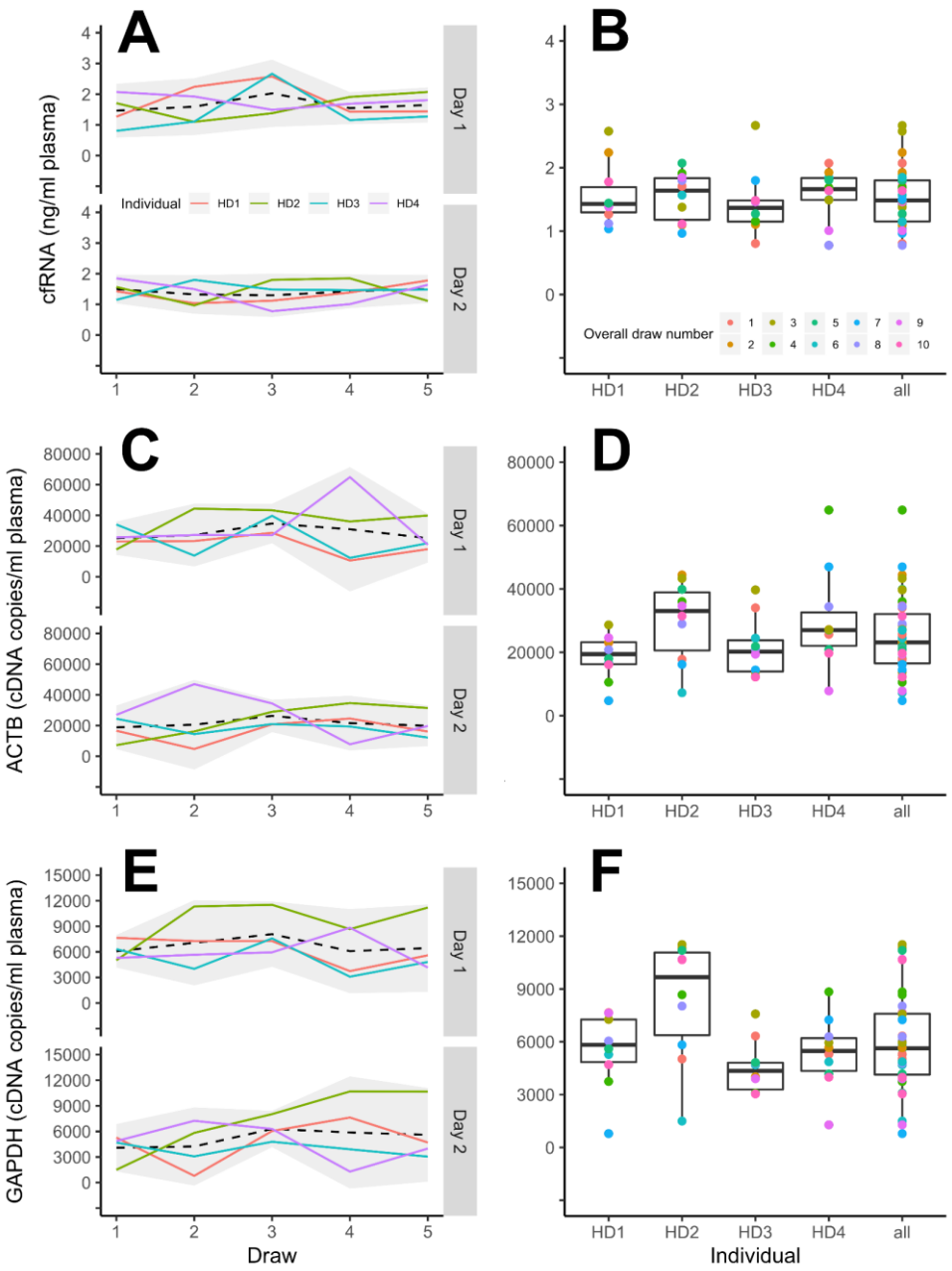


Figure 3.3 | Abundance of plasma-derived cfRNA across the five sampled time points.

CfRNA abundance was measured by Bioanalyzer (A,B) and by ddPCR with probes targetting *ACTB* (C,D) or *GAPDH* (E,F). For left panels, dashed lines represent the average of the four individuals and the shaded area corresponds to 95% confidence intervals. For right panels, the individual data points overlaying boxplots are color coded sequentially starting from the first blood draw of day 1.

ACTB and *GAPDH* are commonly used mRNA normalization genes for liquid biopsy applications [205, 206] despite accumulating evidence suggesting their expression can be highly variable and require situation-specific considerations [207-209]. Currently, there are few reports describing the long-term stability of plasma mRNA expression in individuals. Recent work by Max et al. [202] found no significant variation miRNA plasma/serum profiles following meals and also described interpersonal differences in miRNA abundance that could be stable for up to a year. While we similarly did not find an effect of meals on bulk cfRNA abundance or *ACTB*/*GAPDH* expression, we found significant variation attributed to individual donors for *GAPDH* transcripts. Like our finding with plasma cfDNA abundance, our results suggest that at least some plasma-derived cfRNA transcripts may have baseline expression levels that are specific to the donor; therefore, a thorough understanding of cfRNA normalization transcripts across time and between individuals may be necessary for future cfRNA diagnostic applications.

EVs counts can vary significantly between individuals, but does not vary by draw or day

In order to characterize the abundance of plasma-derived EVs , samples from four individuals across five sampled time points were measured by high-resolution flow cytometry. To standardize the instrument for relative size calibration, we used fluorescent beads of varied diameters (0.1-1um) for approximate and relative sizing of nano-size EVs. To allow comparison and validation of data between the experiments, we fixed 200-nm calibration beads at the start of each experiment to match side scatter (SSC) signal intensity at 10^4 in SSC/FL plot (488-nm excitation; 530/30-nm emission) [210]. With this setup, we

detected FITC-labeled size calibration beads range from 100-nm to 900-nm above noise events (**Supplementary Figure S3.6A**). To characterize relative EV abundance, flow cytometric analysis was performed by gating SSC below 10^4 (using 200-nm bead reference) in SSC/FL plot (**Supplementary Figure S3.6B**). Using uniform fluorescent labeling and appropriate assay controls with EV canonical markers (CD63, CD9, and CD81) as well as platelet marker (CD41), the majority of plasma EVs were detected within this gating (**Supplementary Figure S3.6C, S3.7**). We examined the relative number of EVs stained with canonical exosomal markers CD9 (**Figures 3.4A and 3.4B**), CD81 (**Figures 3.4C and 3.4D**), and CD63 (**Figures 3.4E and 3.4F**), in addition to platelet marker CD41 (**Figures 3.4G and 3.4H**), for each individual and time point. We found no significant differences in relative EV abundancies between the five draws using these markers (**Table 3.2, Supplementary Tables S3.6-S3.10**). When individuals were compared, we observed significant variation between individuals for CD63 and CD81, (**Table 3.2, Supplementary Table S3.7 and S3.8**). Notably in HD3 draw 3, day 1, we observed a spike in counts for CD9 and CD41 that did not appear to follow the trend for other samples and time points (**Figures 3.4A-3.4D**), likely due to outstanding platelet contribution in this draw. The high variation caused by this sample time point may explain why the counts for CD9 and CD41 were not determined to vary significantly between individuals when tested using one-way permutation tests of independence. Non-platelet markers CD81 was not affected by the fluctuation of this draw. When individuals were compared, we observed significant variation between individuals for CD63 and CD81 (CD63 $P < 0.01$, CD81 $P < 0.001$, **Table 3.2**, **Supplementary Tables S3.7 and S3.8**). Individual HD4 had higher level of CD63+ EVs compared to the overall average

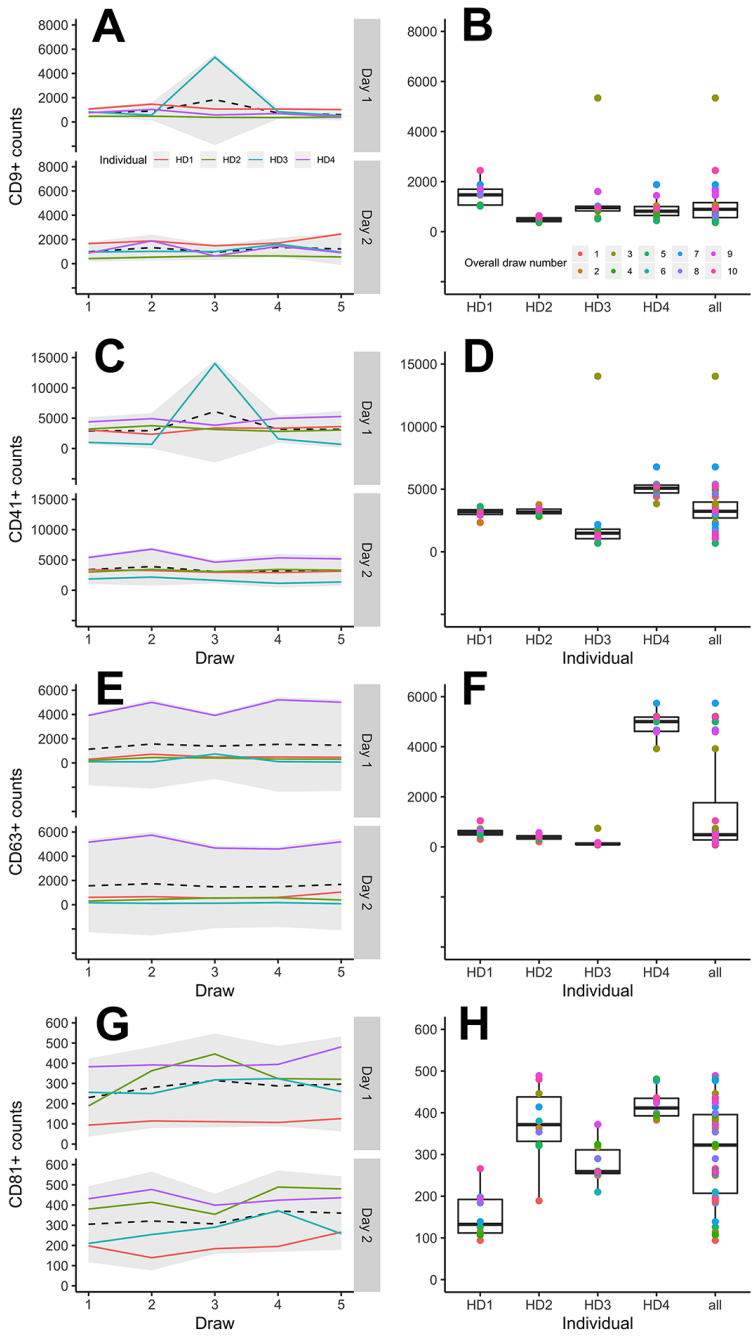


Figure 3.4 | Abundance of plasma-derived EVs across the five sampled time points. EVs were fluorescently labeled using antibodies for CD9 (A,B), CD41 (C,D), CD63 (E,F) or CD81 (G,H) and counted using flow cytometry. For left panels, dashed lines represent the average of the four individuals and the shaded area corresponds to 95% confidence intervals. For right panels, the individual data points overlaying boxplots are color coded sequentially starting from the first blood draw of day 1.

consistently across the draws and days (**Figure 3.4E-3.4F**). Similarly, the level of CD81+ EV of individual HD1 was persistently lower than the overall average throughout all days in two days (**Figure 3.4G-3.4H**). Post-hoc pairwise permutation tests also identified individuals with significantly different CD63 and CD81 EV counts (**Supplementary Tables S3.7-S3.8**). These observations suggest that the primary source of variation in EVs is interpersonal. Conversely, Danielson et al. [211] demonstrated diurnal fluctuation of circulating EVs when characterized by forward and side scattering distributions. Their size gating was placed at > 200 nm gating region, wherein the majority of platelet-derived microvesicles are found, and therefore it is unclear whether their observed distribution of diurnal EV variation was caused in part due from effects of platelet activation. We also note that accumulating evidence suggests the specific blood processing protocol used prior to EV characterization can directly affect final EV composition [196, 212], and thus our study may not be directly comparable to Danielson et al. [211]. Ultimately, our results support a model in which the abundance of canonical human plasma EVs remains stable throughout the daytime and individuals maintain their own distinct EV baselines.

3.5 Conclusions

Liquid-biopsy technology for patient risk-stratification, diagnosis, or disease progression monitoring will likely require biomarker thresholds tailored specifically for each patient. In our pilot cohort, we observed significant interpersonal variation for each of the three analytes examined. Remarkably, distinct baseline levels of cfDNA, cfRNA and EV of each individual were persistent through the draws over time. Future multi-omic studies such as the work presented here, but with larger and more inclusive cohorts, will

be essential for determining the full extent of interpersonal and sample collection variation that may be present across populations.

Funding

This work was supported by funding from the Cancer Early Detection Advanced Research (CEDAR) center at Oregon Health & Science University's Knight Cancer Institute. LN is supported by the National Institute of Child Health and Human Development (1K23 HD091369-01).

Acknowledgements

We thank OCTRI for assistance with obtaining samples and the Knight Cancer Institute Biostatistics Shared Resource for statistics advisement.

Author contributions

Conceptualization: PS, TN, and KJ. Data Curation: JW. Formal Analysis: JW and HK. Funding Acquisition: PS and KJ. Investigation: JW, HK, KJ, LN and TK. Methodology: PS, TN, JW, HK, and KJ. Project Administration: PS and KJ. Resources: PS and KJ. Software: JW. Supervision: TN and PS. Validation: JW and HK. Visualization: JW. Writing – Original Draft Preparation: JW and HK. Writing – Review & Editing: all authors.

Chapter three, in part, is a reprint (with co-author permission) of the material as it appears in the submitted manuscript: “Diurnal stability of cell-free DNA and cell-free RNA in human plasma samples”, Josiah T. Wagner, Hyun Ji Kim, Katie C. Johnson-Camacho, Taylor Kelley, Laura F. Newell, Paul T. Spellman, and Thuy T. M. Ngo, *Scientific Reports* volume 10, Article number: 16456 (2020). The author of this dissertation is the second author of this manuscript.

Chapter IV: Plasma cell-free RNA profiling enables multiclass pan-cancer detection and distinguishes cancer from pre-malignant conditions

4.1 Abstract

Cell-free RNA (cfRNA) in plasma reflects phenotypic alterations of both localized sites of cancer and the systemic host response. Here we report that cfRNA sequencing enables the identification of novel messenger RNA (mRNA) signatures in plasma with the tissue of origin specific to cancer types and pre-cancerous conditions. We sequenced total cfRNA from 66 plasma samples representing three cancer types, two pre-cancerous conditions and healthy donors to explore the diagnostic potential. We identified distinct gene sets and built classification models using the random forest algorithm that could distinguish cancer patients with specific cancer types from premalignant conditions and healthy individuals with high accuracy. Across the four groups that included healthy individuals and patients with lung cancer, liver cancer or multiple myeloma, the cancer types were classified with 96.5% accuracy. 3). Distinction of multiple myeloma from its pre-cancerous monoclonal gammopathy of undetermined significance (MGUS) yielded an accuracy of 90% (17/19). Detection of primary liver cancer from its premalignant condition cirrhosis yielded an accuracy of 100% (12/12). This work lays the foundation for developing low cost assays measuring mRNA transcript levels in plasma using a small panel of genes for identifying cancer types and monitoring pre-malignant disease progression across cancers.

4.2 Introduction

Although recent advances in cancer research offer new methods to treat cancer, early detection of malignancy still constitutes the highest chances of long-term patient survival. For lung cancer, the leading cause of cancer death worldwide, over half of cases

can be cured with existing treatments if detected early while fewer than 5% will survive past 5 years if detected late [213, 214]. Early detection of liver cancer, which has the most rapidly increasing incidence in the United States, would extend 5-year survival rates to 33% with current treatment options. Currently, only 2.4% of metastatic liver cancer patients survive for more than 5 years [58]. Even with a hematologic malignancies like multiple myeloma, 95% of patients are detected when the cancer has already systemically spread, resulting in a decrease of at least 20% in 5-year survival rates compared to when the disease is detected early [215]. Noninvasive, low cost and reliable cancer diagnostic assays could greatly benefit patients by facilitating accessibility to early cancer screening.

For many cancers, there are disease states known to be precursors of malignant disease. For example, multiple myeloma, a cancer of antibody-producing plasma cells, is often preceded by monoclonal gammopathy of undetermined significance (MGUS), which is characterized by lower levels of abnormal antibodies. The prevalence of MGUS is about 3% in the Caucasian population, and the conversion rate from MGUS to multiple myeloma is approximately 1% per year [216, 217]. Hepatocellular carcinoma (HCC), the most common form of liver cancer, is often preceded by liver cirrhosis, which is a degenerative disease characterized by irreversible fibrosis of the liver and is present in 4.5-9.5% of the global population [218-220]. The risk of developing *de novo* HCC in patients with liver cirrhosis ranges between 1-5% per year, depending on the etiology of the cirrhosis [218-224]. Most early cancer detection studies to date have focused on distinguishing cancer from healthy controls, rather than discriminating between cancer and common premalignant conditions. There is an unmet clinical need for a simple blood test that can

identify patients who require further interventions that detect cancer in patients with premalignant conditions during regular surveillance.

With current clinical practices, cancer diagnosis is primarily initiated based upon clinical symptoms that are not generally recognized until tumors are at an advanced stage. Liquid biopsy, a minimally invasive method for the sampling and analysis of analytes in various body fluids, has the potential to improve cancer diagnosis and prognosis [225-228]. Several blood-based analytes have been explored for liquid biopsy utilities in cancer detection such as circulating cells (Circulating Tumor Cells (CTCs), Circulating Hybrid Cells (CHCs), Tumor Associated Macrophages (TAMs)) [229-234], circulating tumor DNA (ctDNA) [235-237], platelets [238-240] and protein panels [241]. However, ctDNA and circulating cells are present at low levels, have very diverse characteristics between patients, and only weakly correlate with phenotypic changes of the tumors [230, 242, 243]. Epigenetic features of ctDNA such as profiles of DNA methylation, 5-hydroxymethylcytosine and ctDNA protected patterns may provide information about the tissue of origin for pan-cancer detection [244-249]. However, these methods are currently expensive due to the requirement of large coverage sequencing. Recent transcriptome analysis of tumor-educated platelets has shown promise for pan-cancer detection, but platelets are fragile, can be easily activated in vitro, and have highly variable characteristics depending on their preparation that make them incompatible with existing clinical blood tests [250]. There is a need for robust liquid biopsy technology that can overcome these challenges in a safe, reliable and cost-effective manner.

Blood flows throughout the body and collects the cell-free RNA (cfRNA) released from cells by active secretion or through cell death including apoptosis and necrosis [251,

252]. Therefore, cell-free transcriptomes from plasma have the potential to reflect the systemic response to growing tumors and information about the tissue of tumor origin specifically by cancer type. Previous work has demonstrated that global cfRNA profiles can reflect temporal abundance changes of organ-specific transcripts, and further analysis through machine learning allows the prediction of pregnancy delivery and preterm birth [253-255]. Here, we explore the potential of cfRNA profiles to distinguish between cancer types and pre-malignant conditions. We sequenced total plasma cfRNA from plasma samples of patients with three cancer types including lung cancer (LuCa), liver cancer (HCC) and multiple myeloma (MM), two pre-cancerous conditions including liver cirrhosis (Cirr) and MGUS, and healthy donors. Feature selection and classification models were built to explore the potential of cfRNA profiles in multiclass cancer detection and differentiating malignant from pre-malignant conditions.

4.3 Materials and Methods

Patient Samples

Blood samples from healthy individuals and patients with monoclonal gammopathy of undetermined significance (MGUS), multiple myeloma, liver cirrhosis, liver cancer, and lung cancer were obtained from Oregon Health and Science University (OHSU) by Knight Cancer Institute Biolibrary and Oregon Clinical and Translational Research Institute (OCTRI). All samples were collected under institutional review board (IRB) approved protocols with informed consent from all participants for research use. Individuals who had no recorded previous history of cancer were considered to be healthy donors.

All lung and liver cancer patients were treatment naïve at the time of blood collection. Treatment naïve is not an excluded criteria for Multiple Myeloma patients. These patients encompass early and late stage cancers. All samples were collected and processed using a uniform protocol by the same staff at Oregon Health and Science University. Samples for analysis were matched between cancer and control groups with respect to age and gender of participants.

Processing of whole blood

For all cohorts, whole blood samples were collected in EDTA-anticoagulated vacutainers. Within 2 h of collection, blood samples were first centrifuged at 1,000g for 10 min at 4°C followed by 15,000g for 10 min at 4°C. Plasma was then stored at -80°C until RNA isolation.

cfRNA isolation

Total RNA purification was performed by using plasma/serum circulating and exosomal RNA purification kit (Norgen Biotek) from 3ml of human plasma according to the manufacturer's protocol. To digest trace amounts of contaminating DNA, RNA was treated with 10X Baseline-ZERO DNase. DNase I treated RNA samples were purified and further concentrated using RNA clean and concentrator-5 (Zymo Research) according to the manufacture's manuals. Final eluted RNA was stored immediately at -80°C.

Library preparation

We prepared stranded RNA-Seq libraries using Clontech SMARTer stranded total RNA-seq kit v2- pico input mammalian (Takara Bio) according to the manufacturer's instructions. For cDNA synthesis, we used option 2 (without fragmentation), starting from highly degraded RNA. Input of 7ul of RNA samples were used to generate cDNA libraries suitable for next-generation sequencing. For addition of adapters and indexes, we employed SMARTer RNA unique dual index kit -96 U. SMARTer RNA unique dual index of each 5' and 3' PCR primer were added to each sample to distinguish pooled libraries from each other. The amplified RNA-seq library was purified by immobilization onto AMPure XP PCR purification system (Beckman Coulter). The library fragments originated from rRNA and mitochondrial rRNA were treated with ZapR v2 and R-Probes according to manufacturer's protocol. For final RNA-seq library amplification, 16 cycles of PCR were performed and final 20 ul was eluted in Tris buffer following amplified RNA-seq library purification. The amplified RNA-seq library was stored at -20°C for sequencing.

Sequencing data processing and quality control

Each sample was sequenced to more than 20 million paired-end reads using an Illumina Nextseq or HiSeq sequencer. Adapter sequences were trimmed using sickle tool [256]. After trimming, the quality of the reads were checked using FastQC (v0.11.7) [257, 258] and RSeQC (v2.6.4) [259]. Reads were aligned to the hg38 human genome using the STAR aligner (v2.5.3a) [115] with two pass mode flag. Duplicated reads were removed using the picard tool (v1.119) [260]. Read counts for each gene were calculated using the htseq-count tool (v0.11.2) [261] in intersection-strict mode. The number of mapped reads

to each gene were normalized to the total number of reads in the whole transcriptome (Reads Per Million - RPM). For each sample, we calculate exon, intron, intergenic fractions and protein coding fractions (CDS exons) using RSeQC [262]. Samples with an exon fraction larger than 0.35 were kept for further analysis.

Identification of cfRNA biomarkers (DESeq and LVQ and GO analysis)

Two independent methods were applied to select cfRNA features for building classification models. Differentiating genes between all pairwise comparisons were identified with the R package DESeq2 (v1.24.0) using the Wald test [263]. The second method for feature selection using the LVQ algorithm built in an R package caret (v6.0-84) - with 10 fold cross validation repeated 3 times [264]. The top 10 most important features were selected by ranking the varImp parameter. Gene Ontology (GO) analysis was implemented on the top differentiating genes from the DESeq2 analysis with padj > 0.01 using the package topGO (v2.37.0) and a Fischer statistical test to measure significant enrichment of each Gene Ontology term [265].

Cancer type classification (LDA and RF)

Two methods were used to build models for classifying cancer types using feature sets identified from pairwise comparison using DESeq2 and LVQ methods. LDA models are built using the R package MASS (v7.3-51.4) [266]. Random Forest models were built using the R package randomForest (v4.6-14) [267].

Statistical consideration (permutation test and leave one out cross validation)

To test if the difference in pairwise comparison between each cancer type and healthy control was specific, a permutation test in which differential expression analysis using DESeq2 package was performed between two groups of randomized samples. For each pair, 500 permutations of random shuffling were performed and the number of differentiating genes with padj < 0.01 were documented for building a histogram, and compared to the number of significant genes (padj < 0.01) for the group with correct labeling. To determine the significance and accuracy of our classification models, we employed the LOOCV method. Briefly, in LOOCV, LDA or RF algorithms classifies each sample based on the training models obtained from all other samples (total number of samples in each pair minus the testing sample). The test was repeated until all individual samples were classified and cross-validated.

4.4 Results

cfRNA profiles distinguish between healthy individuals and those with cancer

We prospectively collected blood samples from a pilot set of 34 cancer patients including 15 LuCa, 10 MM and 8 HCC; 13 pre-malignant conditions including 9 MGUS and 4 Cirr; and 20 age and gender matched healthy donors. Samples were randomly shuffled for RNA extraction, library preparation and sequencing in Illumina flow cells (**Figure 4.1a**). Libraries were sequenced to a mean of 34M raw reads with a range of 28M to 52M (**Supplementary Table S4.1**, **Supplementary Figure S4.1**). After selecting for reads that mapped uniquely to the human genome, the cfRNA libraries had effective read depth of 7M with a range from 2M to 22M. On average, 79% of reads mapped to exons

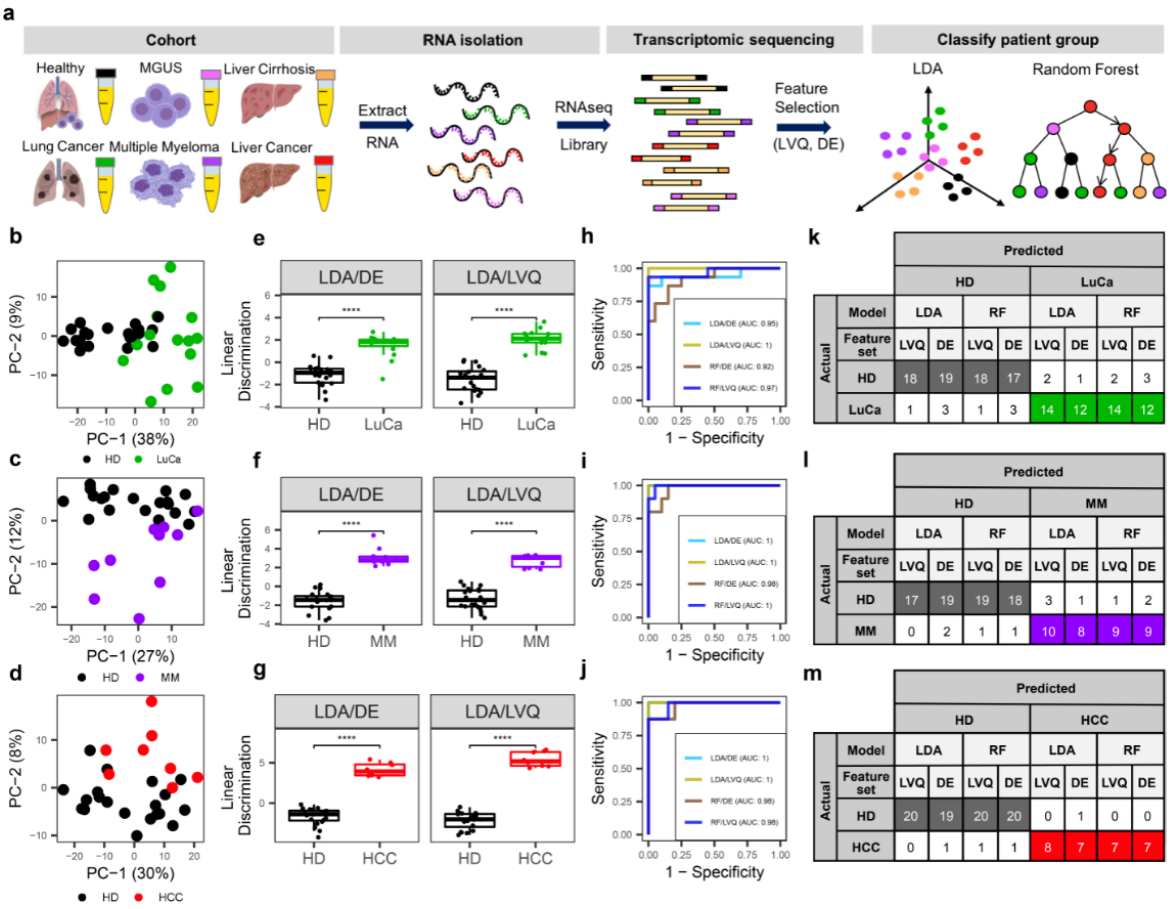


Figure 4.1 | cfRNA profiles distinguish between cancer vs. healthy donors.

(a) Schematic overview of the cfRNA profiling workflow as starting from 3 mL of plasma collected from the patients and healthy donors in EDTA-coated tubes, cfRNA extraction, sequencing, feature selection and classification. (b-d) PCA analysis using top 500 genes with largest variance across healthy and lung cancer samples (b), or multiple myeloma (c) or liver cancer sample (d). (e-g) Linear Discriminant Analysis (LDA) using DE genes with padj < 0.01 and top 10 most important genes identified by LVQ analysis. P-value is derived from Wilcoxon test. (h-j) ROC curves of the two classification models LDA and random Forest (RF) model with two feature sets DE and LVQ. (k-m) LOOCV with the two models LDA and RF with two feature sets DE and LVQ.

(**Supplementary Table S4.1**, **Supplementary Figure S4.2**). A total of 40,226 annotated features were detected with at least 1 mapped read across all samples. The majority of detected RNAs are protein coding with a mean fraction of 81% in the range from 65% to 89% (**Supplementary Table S4.1**, **Supplementary Figure S4.3**). The fraction of reads mapping to exons and the distribution of read depths were uniform across all sample groups (**Supplementary Figure S4.1-S4.2**).

We determined whether cfRNA profiles can distinguish cancer from normal controls for all 3 tested cancer types: lung cancer, liver cancer and multiple myeloma. Unbiased Principal Component Analysis (PCA) using the top 500 genes with the largest variance across all samples through pairwise comparison showed separation of LuCa, HCC and MM cfRNA profiles from that of healthy donors (**Figure 4.1b-4.1d**). Differential expression (DE) analysis of pairwise comparison between individual cancer types with respect to healthy donors using DEseq2 yielded 1864, 110, and 12 differentiating genes (adjusted p-value < 0.01) for LuCa, MM and HCC, respectively (**Supplementary Figure S4.4**). To confirm the significance of our differential expression results for each pairwise comparison of cancer to healthy donors, we performed a permutation test in which differential expression analysis between two groups of randomized samples was compared. Permutations of random sample shuffling in each pair with 500 rounds resulted in zero significant differentiating genes (padj < 0.01) in more than 95%, 95% and 94% of permutations for each pair comparing LuCa, MM, and HCC to healthy donors, respectively (**Supplementary Figure S4.5**). GO analysis revealed that up-regulated genes in LuCa were enriched for neutrophil activation and aggregation (**Supplementary Figure S4.6a**). In MM, oxygen and gas transport were the most enriched processes in the up-regulated gene

set (**Supplementary Figure S4.6b**). In HCC, the up-regulated gene set was enriched for plasminogen activation (**Supplementary Figure S4.6c**). This data collectively indicates the separation of cfRNA profiles in LuCa, HCC and MM compared to healthy donors.

To explore the potential of cell-free RNA for cancer detection, we applied Linear Discriminant Analysis (LDA) and a Random Forest (RF) algorithm to find combinations of discriminating genes to separate cancer from healthy individuals. Two independent methods were used to identify specific input gene lists for the classifying algorithms. First, discriminating genes using DESeq2 analysis with False Discovery Rate (FDR/adjusted p-value) < 0.01 were used as one feature set (DE gene set). Second, we implemented the learning vector quantization (LVQ) method to find the most important features that distinguish the two groups and selected the top 10 as another feature set (LVQ gene set). The linear combination for each gene set by LDA showed significant separation between LuCa, HCC and MM from healthy donors with p-value of $5.6x10^{-7}$ and $6.2x10^{-10}$, $6.7x10^{-8}$ and $6.7x10^{-10}$ and $6.4x10^{-7}$ and $6.4x10^{-7}$ using the DE and top 10 LVQ gene sets, respectively (**Figure 4.1e-4.1g**). We further employed the Random Forest (RF) method to develop orthogonal classification models. The area under the receiver operating characteristic (ROC) curve (AUC) is higher than 0.92 in both LDA and RF models for both DE and LVQ feature sets of all three cancer types (**Figure 4.1h-4.1j**).

To evaluate the significance and accuracy of our classification models, we employed the leave-one-out cross validation (LOOCV) method. Briefly, in LOOCV, one sample was iteratively removed for testing, with the remaining samples used for training by the LDA or RF algorithms to create a classifying model. LDA or RF algorithms classified each left out sample based on these training models. The test was repeated until

all individual samples were classified and cross-validated. Both LDA and RF algorithms were trained on the described DE and LVQ gene sets, resulting in four classification models (**Figure 4.1k-4.1m**). Classifying LuCa from healthy donors yielded accuracies of 91% (32/35), 88% (31/35) when using the LDA method, and 91% (32/35), 83% (29/35) when using the RF method with LVQ and DE feature sets. The overall successful prediction rates for differentiating MM from healthy donors are greater than 90% (27/30) for all four combinations. HCC were correctly differentiated from healthy donors with accuracies of 100% (28/28) and 93% (26/28) when using the LDA method and 96% (27/28) and 96% (27/28) when using the RF method with LVQ and DE feature sets. Overall, the LOOCV test confirmed that the biomarker sets determined by DESeq2 and LVQ methods combined with our classification models using LDA and RF algorithms are statistically significant.

cfRNA profiles enable multiclass cancer detection

The feature sets identified by both DESeq2 and LVQ methods are cancer- and organ site- specific (LuCa, HCC and MM) compared to healthy donors (**Figure 4.2a-4.2c).** For example, the top 5 most significant genes from the LVQ analysis discriminates LuCa from healthy donors with a p-value of less than 10^{-7} but when comparing other cancer groups (HCC and MM) to healthy donors the p-value is not significant (**Figure 4.2a**). Similarly, the gene sets for MM (**Figure 4.2b**) and HCC (**Figure 4.2c**) are significantly different for these groups compared to healthy donors with a p-value of less than 10^{-5}, whereas cross comparison of the other cancer types to healthy donors gave non-significant p-values. Therefore, we attempted to develop multi-class classification models to explore if cfRNA profiles could enable pan-cancer detection. The top 5 most important in 6 pairwise

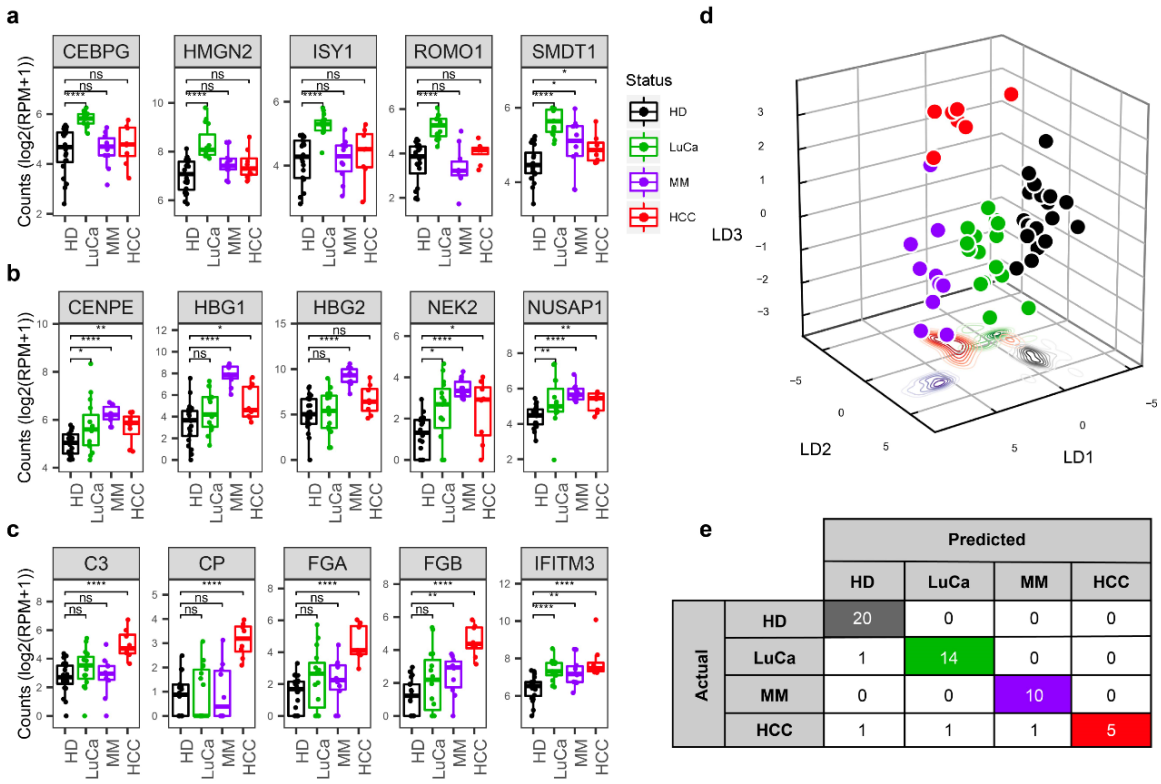


Figure 4.2 | cfRNA profiles enable classification of pan-cancers.

(a-c) Box plots of representative top 5 most important genes resulted from LVQ analysis for three pairs: multiple myeloma (**a**), lung cancer (**b**) and liver cancer (**c**) versus healthy. (**d**) Multiclass discrimination using LDA classification with combination of 5 most important genes identified LVQ analysis from pairwise comparison of six pairs between lung, liver, multiple myeloma and healthy. (**e**) LOOCV using RF algorithm.

comparisons between each cancer type, healthy controls and other cancer types from the LVQ analysis are combined as a feature set of 30 genes for multi-class classification. LDA using the combination of genes in this feature set displayed clear separation between three cancer types and healthy donors (**Figure 4.2d**). The RF algorithm classified across cancer types with high accuracy as assessed by LOOCV (**Figure 4.2e**). The RF model accurately

categorized 20/20 healthy donors, 14/15 LuCa, 10/10 MM and 5/8 HCC samples (92.5% accuracy overall). Cohen's kappa coefficient showed approximately 84% agreement between the RF model and the actual diagnosis. This result demonstrates the feasibility of our cfRNA platform to detect not only the presence of cancer, but also the specificity of the cancer tissue of origin.

cfRNA profiles distinguish multiple myeloma from its premalignant condition, MGUS, and MGUS from healthy

We examined if cfRNA profiles were able to recapitulate the transition from a pre-cancerous condition to a cancerous one, and distinguish between them. We chose to test our hypothesis on multiple myeloma (MM) as it has a well-defined pre-cancerous condition, MGUS. The top ten most significant genes that discriminate MM from HD as identified by LVQ displayed a gradual transition in cfRNA level from the HD through MGUS to MM (**Figure 4.3a**). Among these ten most significant genes, nine genes (AIDA, CA1, CENPE, CPOX, EPB42, HBG1, HBG2, NEK2 and NUSAP1) are expressed higher in B cells and bone marrow compared to other tissue and cell types [268]. Three out of the ten most important genes resulting from the LVQ analysis are related to cell cycle processes: Centromere protein E (CENPE), a kinesin-like motor protein that accumulates in the G2 phase of the cell cycle and is highly expressed in bone marrow [269, 270]; Serine/threonine-protein kinase (NEK2), which is involved in mitotic regulation [270, 271]; and Nucleolar and spindle associated protein 1 (NUSAP1), a nucleolar-spindle-associated protein that plays a role in spindle microtubule organization [272].

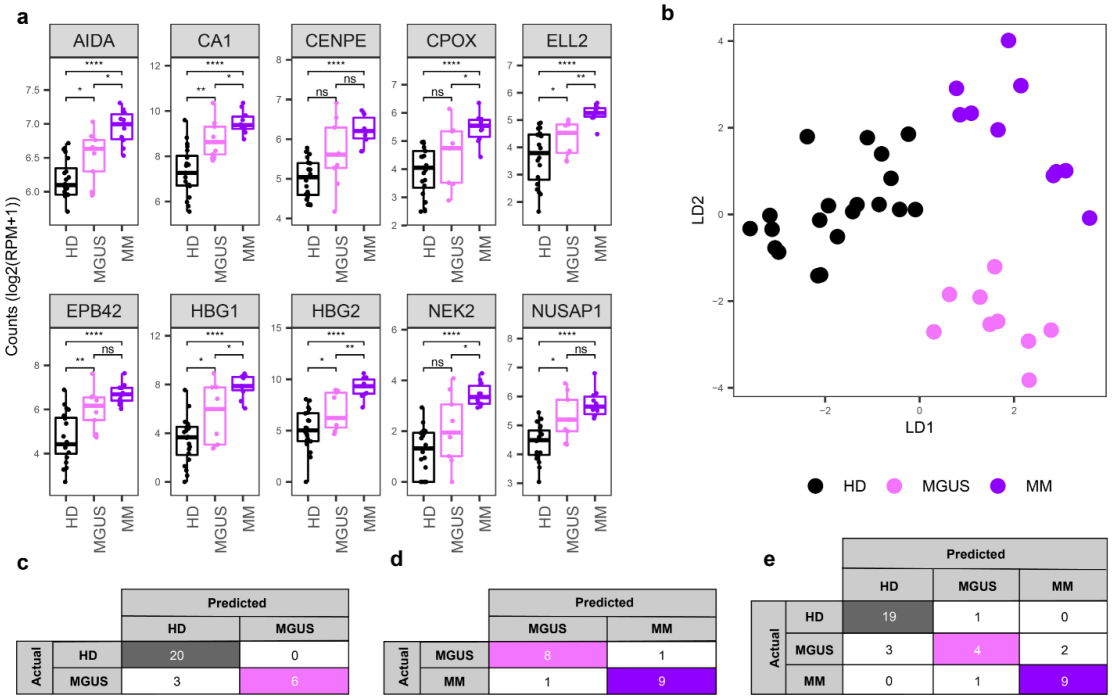


Figure 4.3 | cfRNA profiles distinguish between healthy, MGUS and multiple myeloma donors:

(a) Box plots of representative top 10 most significant genes resulted from learning vector quantization analysis for multiple myeloma versus healthy. (b) LDA plot using 10 genes from pairwise analysis across healthy - MGUS and healthy - multiple myeloma samples using the learning vector quantization method. (c-e) LOOCV using 2 models (LDA and RF) with top 10 lvq genes to discriminate MGUS and healthy (c), multiple myeloma vs MGUS (d) and three groups healthy, MGUS and multiple myeloma (e).

An LDA plot using a combination of the top 10 LVQ genes from pairwise comparisons MM- healthy donor, and MGUS- healthy donor displayed the separation of all three groups (**Figure 4.3b**). A RF model using the top 10 most important LVQ genes from MGUS- healthy donor pairwise comparison yielded an accuracy of 88.6% (20/20 healthy donors and 6/9 MGUS patients) (**Figure 4.3c**). Classification of MM from MGUS yielded an accuracy of 89.5% (8/9 MGUS and 9/10 MM) using LOOCV with the RF

classification method using the top 10 most important genes from LVQ analysis of MM versus HD comparison as a feature set (**Figure 4.3d**). The 3-group classification resulted in an accuracy of 82% (19/20 healthy, 4/9 MGUS and 9/10 MM) defined by LOOCV using the RF method with the feature set composed of the combination of the top 10 LVQ genes from the comparison MM versus HD and MGUS versus HD (**Figure 4.3e**).

cfRNA profiles distinguish liver cancer from its pre-malignant condition, cirrhosis, and cirrhosis from healthy

Next, we asked if we could distinguish between a solid tumor such as hepatocellular carcinoma (HCC) from its pre-cancerous condition, liver cirrhosis (Cirr). Among the top ten most important genes that discriminate HCC from HD identified by the LVQ analysis, five genes also significantly differentiate HCC from Cirr (**Figure 4.4a**). Interestingly, 8 out of the top 10 genes are expressed specifically in the liver and the corresponding proteins are secreted to the plasma [268]. Apolipoprotein E (APOE) binds to a specific liver and peripheral cell receptor and is essential for normal catabolism of triglyceride-rich lipoprotein constituents [273]. Complement C3 (C3) is synthesized in the liver and is involved in both innate and adaptive immune responses [274]. Ceruloplasmin (CP) is a secreted plasma metalloprotein from the liver that binds copper and is involved in the peroxidation of Fe (II) transferrin to Fe (III) transferrin [275]. 24-dehydrocholesterol reductase DHCR24 catalyzes the reduction of sterol intermediates [276]. Fibrinogen Alpha Chain (FGA), fibrinogen Beta Chain (FGB) and Fibrinogen Gamma Chain (FGG) encode the coagulation factor fibrinogen, which is a component of blood clotting [277]. Histidine Rich Glycoprotein (HRG) is a plasma glycoprotein that binds heparin sulfate on the surface of the liver, lung, kidney and heart endothelial cells [278].

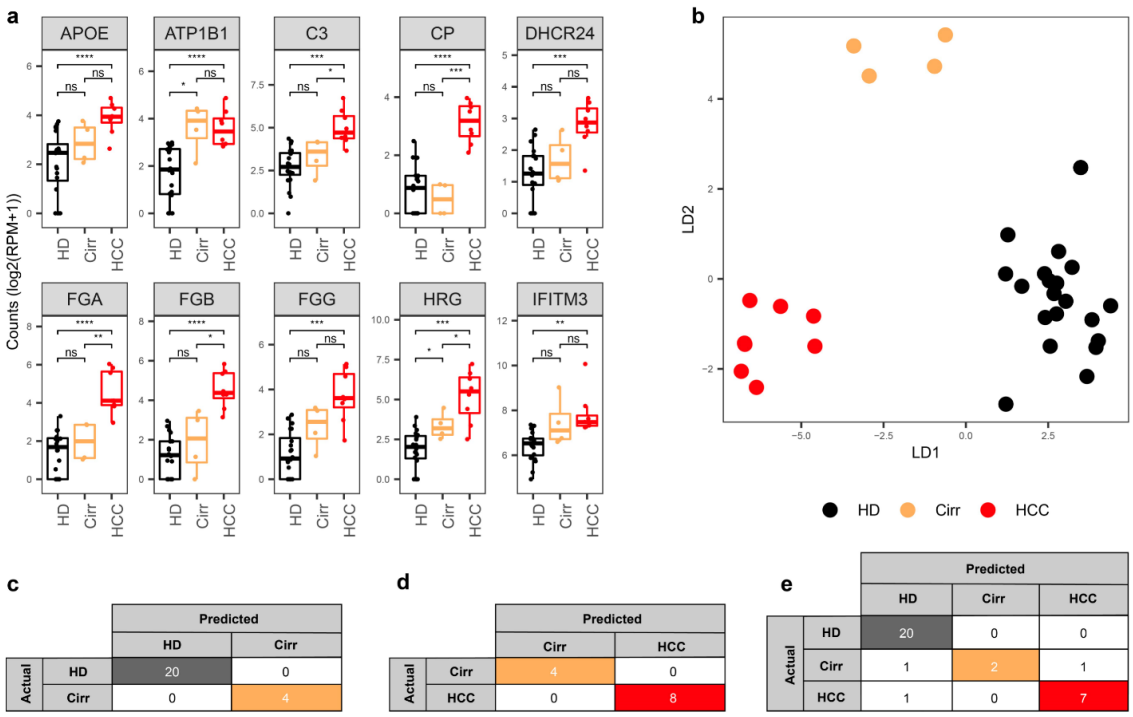


Figure 4.4 | cfRNA profiles distinguish between healthy, liver cirrhosis and liver cancer donors:

(a) Box plots of representative top 10 most significant genes resulted from learning vector quantization analysis for liver cancer versus healthy. (**b**) LDA plot using top 10 genes from each pairwise analysis between healthy - liver cirrhosis and healthy - liver cancer samples using the learning vector quantization. (**c**-**e**) LOOCV using 2 models (LDA and RF) with top 10 lvq genes to discriminate liver cirrhosis and healthy (**c**), liver cancer vs cirrhosis (**d**) and three groups healthy, liver cirrhosis and liver cancer (**e**).

Current practices for HCC surveillance include screening on Cirr patients using imaging techniques, such as ultrasound and MRI. These methods are expensive and can have limited accessibility [218]. In addition, detection of Cirr is mostly based on clinical symptoms which are often displayed at late stages of the disease [279]. Therefore, easy-to-use, reliable and specific biomarkers with accompanying prediction models are needed to improve detection of both HCC and Cirr. We explored the potential of cfRNA to

distinguish HCC from Cirr and Cirr from healthy individuals. An LDA plot using the feature set comprised of a combination of the top 10 LVQ genes identified for the pairwise comparisons of HCC- healthy donor and Cirr- healthy donor, shows a distinct separation between these groups (**Figure 4.4b**). RF methods using the top 10 important genes from Cirr- healthy donor pairwise comparisons yielded 100% accuracy in classifying Cirr from healthy donor samples using LOOCV (**Figure 4.4c**). Classification of HCC from Cirr also yielded 100% accuracy using LOOCV with RF (**Figure 4.4d**). We further attempted to classify three classes including HD, Cirr and HCC in one model. The 3-group classification resulted in 90.6% accuracy using LOOCV with RF (**Figure 4.4e**).

4.5 Discussion

We sequenced cfRNA from patients with three types of cancer (LuCa, HCC and MM), two pre-cancerous conditions (Cirr and MGUS) and healthy donors. All three cancer types can be distinguished using their cfRNA profiles which allowed the use of machine learning algorithms trained with a panel of cell-free RNA transcripts to accurately classify the three cancer types. To differentiate each cancer type from healthy individuals, the combination of ten genes identified by learning vector quantization (LVQ) analysis in each pairwise comparison yields similar accuracy compared to the use of a larger set of differentiating genes evaluated by leave one out cross validation. Two classification models built on linear discriminant analysis (LDA) and the random forest (RF) algorithm resulted in similar classification performance in each pairwise comparison of cancer to healthy donors. To distinguish each group in a multiclass panel including LuCa, HCC, MM and healthy donors, a panel of 30 genes gave a classification accuracy of 92.5% using a RF

model. The use of a small gene panel potentially enables a cost-effective assay for pan-cancer detection that could be performed in a doctor's office.

To date, most investigations into the potential of blood-based methods for cancer detection have only focused on distinguishing cancers from healthy controls [228, 235, 238, 239, 241, 249]. However, many cancer types have a long period of precursor states such as MGUS for MM and Cirr for HCC. Here, we report that cfRNA profiles can recapitulate the transition from a pre-cancerous condition to cancer. We therefore propose that cfRNA panels containing a small number of genes may distinguish cancers from pre-malignant conditions and precursors from healthy individuals. This development can potentially enable a cost-effective screening strategy for early cancer detection during routine exams in high-risk pre-malignant patients within the general population.

Lung, liver and bone marrow have been reported to contribute heavily to the abundance of cell-free nucleic acids in plasma [253]. This may explain the source of cfRNA biomarkers found in these three cancer types. In LuCa, upregulated genes are enriched for immunity markers related to neutrophil activation, which might partially reflect inflammation of the tumor microenvironment. In HCC, nine out of the top ten genes used in the classification model are specifically synthesized in the liver and encode secreted proteins found in blood that mediate plasminogen activation and fibrinolysis processes. In MM, seven out of ten genes among the most important cfRNA biomarkers have relatively high expression in B cells and bone marrow compared to other tissue and cell types and are related to cell cycle processes. These findings indicate that the cfRNA biomarkers identified likely originate from the tissue of origin of the tumor.

Our pilot study has important limitations. The discovery was performed using a small sample cohort and does not contain large scale independent sample sets. Further they do not represent the distribution of cancer and precursor lesions in the overall population. In addition, we have not fully characterized the stability of cell-free RNA and the biological origin of the identified cfRNA biomarkers. Before the tests developed from this work can be clinically applied, large-scale clinical studies will be required to validate the potential of cfRNA and to build robust classification models.

4.6 Conclusions

In summary, we report the first proof-of-principle that global profiling of cell-free mRNA has the potential to enable a multiclass cancer detection. This work lays the foundation for developing inexpensive assays that measure transcript levels of mRNA in plasma for a small panel of genes that can differentiate pan-cancer from pre-malignant conditions and otherwise healthy donors. Organ-specific mRNA transcripts were identified as biomarkers that indicate the tissue of origin for the tumor. These cell-free plasma RNA profiles could be readily combined with other nucleic acids-based and protein-based approaches for potentially increased diagnostic sensitivity and specificity.

Acknowledgements

This work was supported by the Cancer Early Detection Advanced Research Center (CEDAR), Knight Cancer Institute, Oregon Health and Science University (CEDAR 3250918), Cancer Research UK/OHSU Project Award (C63763/A27122), and the OCTRI

CTSA grant (UL1TR000128). We would like to acknowledge the CEDAR repository and the Biolibrary for helping with sample collection and processing.

Author contributions

T.T.M.N designed and supervised all aspects of the study; T.T.M.N wrote the manuscript; All authors contributed to and edited the manuscript; H.J.K, J. T. W. and T.N. designed experiments and carried out RNA extraction and library preparations; T.T.M.N., P.A. and B.R. processed the sequencing data; T.T.M.N., P.A. and B.R. developed statistical tools and analyzed the data; F.C., P.S., R.F.T, S.N. contributed to the study design, data interpretation and the manuscript.

Competing interests

An invention disclosure based on this work was submitted to the patent office at Oregon Health and Science University in preparation for a patent application. The authors declare no other competing interests.

Code availability

In-house scripts used in this manuscript, which includes data processing, downstream analysis and the scripts used to generate figures publicly available on github: https://github.com/ohsu-cedar-comp-hub/cfRNA-seq-pipeline-Ngo-manuscript-2019

Chapter four is a draft of the manuscript entitled: “Plasma cell-free RNA profiling enables multiclass pan-cancer detection and distinguishes cancer from pre-malignant conditions”, Hyun Ji Kim, Breeshey Roskams-Hieter, Pavana Anur, Josiah T. Wagner, Fehmi Civitci, Paul Spellman, Reid F. Thompson, Willscott E. Naugler, and Thuy T. M. Ngo, In preparation for resubmission (2021). The author of this dissertation is the first author of this manuscript.

Chapter V: Selective packaging of extracellular vesicles RNA association with cancer progression

5.1 Abstract

Cell-free messenger RNA (cf-mRNA) circulates in the bloodstream and has shown potential to be developed as blood-based biomarkers to distinguish cancer and high-risk groups. However, there is limited understanding of what the potential carriers of cf-mRNA are in human plasma, and which cf-mRNAs may exist in either in extracellular vesicle or non-vesicle associated fractions. Additionally, how cf-mRNA packaging profiles are associated with cancer progression is not well understood. Here, we used size exclusion chromatography to characterize cf-mRNA in human plasma samples from three different cancer types (lung cancer, liver cancer, and multiple myeloma), two high-risk groups (monoclonal gammopathy of undetermined significance and liver cirrhosis), and healthy controls. By separating and sequencing potential carriers of cf-mRNA in their respective extracellular vesicle and non-vesicle associated fraction, we found that the majority of cf-mRNA was enriched in extracellular vesicles. Furthermore, cf-mRNA was also protected in membrane-bound vesicles, revealing a remarkable stability in RNase-rich environments. To reveal cf-mRNA packaging associated with cancer progression, cf-mRNA transcripts levels between cancer and high-risk or healthy cohorts were identified in each fraction. Our results suggest that cf-mRNA is not only predominately associated with the EV enriched fraction, but also suggest cancer-distinguishing cf-mRNA transcripts are selectively packaged within human plasma.

5.2 Introduction

Cell-free RNA (cfRNA), also called extracellular RNA, in the blood has shown great potential as a biomarker of disease [280-283]. Cell-free RNA contains both protein-

coding and non-coding RNA. Prior studies have focused primarily on cell-free miRNAs (cf-miRNAs), which are approximately 22 nucleotides in length, and frequently dysregulated in cancer [68]. Although cell-free RNA has been considered more fragile than cell-free DNA because of high RNase levels in blood [140], several studies have demonstrated that miRNAs are present in blood in a remarkably stable form [39, 68]. This stability is likely the reason why the majority of cfRNA studies exploring their role as cancer biomarkers have focused on characterizing cf-miRNA [39, 68]. Although cell-free messenger RNAs (cf-mRNA) are potentially more fragmented and less abundant than cf-miRNAs, recent reports have demonstrated cf-mRNAs carry disease specific signatures that can be exploited as non-invasive biomarkers [70, 284] & [cfRNA paper]. In our previous study, we have also shown increased levels of cf-mRNA correlated with the diagnosis of cancer and with disease progression [cfRNA paper]. Although there are reports of cf-mRNA as ideal non-invasive biomarker, a comprehensive study of how cf-mRNA is protected within endogeneous RNase enriched plasma remains unknown. Understanding the stability of cf-mRNA is an important prerequisite for utility as a blood based biomarker.

This point led us to speculate that cf-mRNA may have been protected in extracellular vesicles [139, 285-288]. Extracellular vesicles (EVs) are a collective term for various vesicles separated based on their size and biogenesis [30]. EVs which are > 100 nm have been categorized as medium EVs, while EVs < 100 nm have been categorized as small EVs [152]. Recently, exomeres, defined as < 50 nm non-membranous nanoparticles, were identified [30, 288]. EVs are known to harbor variable molecular cargoes including nucleic acids and proteins associated with their cells of origin, providing a great potential

for diagnostics and prognostics [280, 289, 290]. The discovery of EVs containing cfRNA has sparked considerable interest in understanding the role of these vesicles in intercellular communication and the potential clinical applications [139, 287]. However, little is known regarding i) whether EVs or protein complexes are the major carrier of cell-free mRNA in complex biofluids such as plasma, and ii) whether circulating cfRNA associated with EVs (EV-RNA) can distinguish cancer patients from healthy controls. Investigation of cfRNA carriers which provide form of stability in plasma and molecular composition of disease specific EV cfRNA may reveal the basic biology, function and clinical translation potential.

Extracellular miRNA have shown to either enclosed within extracellular vesicles or associated with protein complexes [39, 63, 291-293]. Since RNA binding proteins (e.g. Argonaut 2) are known to contain miRNA binding sites, prior studies have investigated whether these complexes are the major form of miRNA carriers [39, 286, 294]. For the first time, presence of miRNA in circulation were found to be primarily associated with the Argonaut 2, which adds even further to the complexity of several potential cf-RNA carries [39]. Therefore, the exRNA atlas, a data repository of the NIH extracellular RNA communication consortium, analyzed cell-free miRNA cargo types from various human biofluids covering 23 healthy conditions across 19 different studies [138]. Understanding type of encapsulation is important as they are related to their roles, functions, and even their destinations [287, 295]. Murillo et al. integrated computational analysis revealed major types of non-coding RNA carriers: extracellular vesicles, RNA binding proteins, or as part of lipoprotein particles, mostly HDL [138, 296]. Despite progress made in

understanding non-coding RNA carries, it remains unclear how cf-mRNAs are associated with different types of carriers.

In the context of cancer progression, oncogenes such as KRAS have been shown to influence the selective packaging of genetic materials into vesicles in cell culture media [63]. Post-transcriptional regulation of Argonaut 2 (Ago2) resulted in dysregulation of miRNAs into exosomes. These exosomal miRNA were demonstrated to potentially affect recipient cell phenotypes, including gene expression, and even cancer invasiveness [63]. Similarly, EV-mediated miRNA delivery to recipient cells has been suggested to form pre-metastatic niche and promote tumorigenesis [291, 297]. Ongoing efforts to understand how extracellular RNA, including long noncoding RNA, is regulated by oncogenic signaling into EVs have been demonstrated in cell culture [139]. Despite recent works demonstrating selective RNA packaging in relation to cancer in cells, identifying mRNA content of EVs in human plasma has remained challenging.

In this study, we utilized size exclusion chromatography to characterize vesicle associated RNA and non-vesicular carrier such as lipoprotein or RNA binding protein complexes. We fractionated plasma into three fractions associated with medium EVs, small EVs, and exomeres, and three non-EV fractions associated with early-, center-, and late-protein elution peaks. These fractions were confirmed by physical characterization of EVs and protein. Subsequently, we extracted and sequenced RNA content of the six fractions of five healthy donors, five lung cancer, five liver cancer, five multiple myeloma, four liver cirrhosis and four monoclonal gammopathy of undetermined significance patients. In total, RNA sequencing was performed on 168 samples and the majority of detected total RNA were protein coding transcripts. Through implementation of novel RNA normalization

across plasma fractions with an external RNA control spike-in, we found that cf-mRNA is predominately found in EV fractions. Furthermore, we identified sets of genes within each fraction whose expression is altered in cancer and high-risk patients compared to healthy donors.

5.3 Materials and Methods

Clinical sample and preparation of plasma

Blood samples from healthy individuals and patients with monoclonal gammopathy of undetermined significance (MGUS), multiple myeloma, liver cirrhosis, liver cancer, and lung cancer were obtained from Oregon Health and Science University (OHSU) by Knight Cancer Institute Biolibrary and Oregon Clinical and Translational Research Institute (OCTRI). All samples were collected under institutional review board (IRB) approved protocols with informed consent from all participants for research use. Healthy donors were individuals who had no recorded previous history of cancer. All lung and liver cancer patients were treatment naïve at the time of blood collection. Treatment naïve was not an excluded criteria for Multiple Myeloma patients. All samples were collected and processed using a uniform protocol by the same staff at Oregon Health and Science University. Samples for analysis were matched between cancer and control groups with respect to age and gender of participants. Whole blood was collected from all clinical samples in 10 mL in K2EDTA tubes (BD Vacutainer, Becton Dickinson, 36643). Tubes were transported vertically at room temperature before processing. Within 1 hour of blood withdrawal, plasma was prepared by centrifugation (Eppendorf 5810-R centrifuge, S-4-104 Rotor, Eppendorf) by double spin condition. 10 ml of whole blood was first spun at 1,000 × g for

10 minutes at 4°C. The supernatant was collected 10 mm above the buffy coat. The second centrifugation was done at 15,000 x g for 10 minutes at 4°C. Aliquots of platelet-depleted plasma were transferred to 1.5 mL microcentrifuge tubes (VWR, 89126-714) and stored immediately at −80 °C.

Plasma fractionation using size exclusion chromatography

Size exclusion chromatography was conducted using commercially-available qEV2 SEC column (Izon Science Ltd, New Zealand) according to the manufacturer's instructions. In brief, the column was equilibrated with 0.1 µm filtered D-PBS without calcium and magnesium (Thermo Fisher Scientific, 14190250) at room temperature. 2 ml plasma was loaded on column, and 14 ml of void volume was discarded. The exact 4 ml of 6 fractions (FR14, FR58, FR912, FR1619, FR2326, and FR3033) of SEC column were collected in 50 ml of canonical tube (Falcon) on ice and was immediately followed by RNA extraction.

Size measurement of EVs using qNano and dynamic light scattering

Concentration of particles in isolated EV fraction from size-exclusion chromatography was measured using tunable resistive pulse sensing by qNano (Izon, Cambridege, MA, USA) following instruction manuals. The calibration particles (Izon, CPC100) and EV fraction collected from size-exclusion chromatography was placed in nanopore (Izon, NP150, A37355). Particle concentration was determined by Izon software. Size distribution of particles eluted from size-exclusion chromatography was measured using Zetasizer nanoseries instrument (Malvern Nano zeta sizer).

EV size distribution measurement using transmission electron microscopy

Ultrathin carbon film on lacey carbon support with 400 mesh on copper (Ted Pella, 01824) was glow discharged for 30 seconds using PELCO easiGlow glow discharger (Ted Pella). Isolated EV samples were put on charged grids for 1 min, washed for 30 seconds with MilliQ water, and fixed with 1% uranyl acetate for 30 seconds. Grids with stained samples were air dried at least 30 minutes before imaging. Prepared samples were imaged at 120 kV using FEI Tecnai™ Spirit TEM system. FEI- Tecnai™ Spirit TEM system was interfaced to a bottom mounted Eagle™ 2K TEM CCD multiscan camera and to a NanoSprint12S-B cMOS camera from Advanced Microscopy Techniques (AMT) fast side mounted TEM CCD Camera. Images were collected at 8,000-80,000x magnification under 1-2μm defocus. Images were acquired as 2048 × 2048 pixel, 16-bit gray scale files using the FEI's TEM Imaging & Analysis (TIA) interface on an Eagle™ 2K CCD multiscan camera.

Immunoprecipitation and western blotting

For immunoprecipitation, 200 μL of Magna Bind goat anti-mouse IgG Magnetic Bead slurry (Thermo Scientific, PI21354) were washed with PBS solution and incubated with 10 μg of mouse monoclonal anti-Ago2 (Abcam, ab57113), anti-CD9 antibody (Abcam, ab58989), anti-Apolipoprotein (Santa Cruz Biotechnology, sc-376818) or mouse normal IgG (Santa Cruz Biotechnology, sc-2025) antibodies for 2 h at 4 °C. To account for smaller volumes, the exact 4 ml of 6 fractions (FR14, FR58, FR912, FR1619, FR2326, and FR3033) of SEC column were concentrated by ultracentrifugation at 150,000 g x 6 hours.

The resulting pellet was lysed in 200 µL of IP lysis buffer (Thermofisher Scientific, 87787) supplemented with a halt protease inhibitor cocktail (Thermofisher, 78430). Total of 200 µL of IP lysed samples were mixed with 200 µL of PBS. The preincubated beads and antibody were then added to the 400 µL of sample and incubated overnight at 4 °C. Beads were washed three times with 1% Nonidet P-40 buffer (Sigma Aldrich, 11332473001) and then eluted in 20 µl of NuPage LDS/reducing agent mix and incubated for 10 min at 70 °C to elute the sample. Samples eluted off from the beads were used for western blotting. Western blot was run using Bolt 4-12% Bis-Tris Plus gel (Life technologies, NW04122) and transferred onto PVDF membrane (Thermofisher scientific, LC2002). The membrane was blocked with 1X TBST containing 5% milk and incubated with primary antibodies overnight at 4°C (Sigma Aldrich, M7409-5BTL). The anti-Argonaut-2 antibody (Abcam, ab32381), anti-CD9 (Abcam, ab223052), and anti-apolipoprotein A1 (Abcam, ab64308) were used. After washing with 1X TBST, membrane was incubated with horseradish peroxidase conjugated anti-rabbit secondary antibodies (Cell Signaling, 7074) and washed again to remove unbound antibody. Bound antibodies were detected with supersignal west pico plus chemiluminescent substrate (Thermofisher, 34577).

RNA extraction from fractionated plasma

RNA was extracted from 4mL of fractionated plasma using plasma/serum circulating and exosomal RNA purification kit (Norgen Biotek, 42800) according to the manufacturer's protocol with some modifications. After fractionated plasma samples were lysed at 60°C for 10 min and mixed with ethanol, 10µl of 10^6 diluted ERCC RNA spike-in control mix (Thermofisher, 4456740) was added on ice as an external RNA control for

normalization. ERCC spiked in samples were followed by centrifugation at 1,000 RPM for 2 min. At that point, the manufacturer's protocol was followed and RNA was eluted in 100µl. To digest trace amounts of contaminating DNA, RNA was treated with 10X Baseline-ZERO DNase. DNase I treated RNA samples were purified and further concentrated using RNA clean and concentrator-5 (Zymo Research, R1014) according to the manufacture's manuals. Final eluted RNA was aliquoted and stored at -80°C immediately.

Library preparation

We prepared stranded RNA-Seq libraries using Clontech SMARTer stranded total RNA-seq kit v2- pico input mammalian (Takara Bio, 634414) according to the manufacturer's instructions. For cDNA synthesis, we used option 2 (without fragmentation), starting from highly degraded RNA. Input of 7ul of RNA samples were used to generate cDNA libraries suitable for next-generation sequencing. For addition of adapters and indexes, we employed SMARTer RNA unique dual index kit -96 U (Takara Bio, 634452). SMARTer RNA unique dual index of each 5' and 3' PCR primer were added to each sample to distinguish pooled libraries from each other. The amplified RNA-seq library was purified by immobilization onto AMPure XP PCR purification system (Beckman Coulter, A63881). The library fragments originated from rRNA and mitochondrial rRNA were treated with ZapR v2 and R-Probes according to manufacturer's protocol. For final RNA-seq library amplification, 16 cycles of PCR were performed and final 20 ul was eluted in Tris buffer following amplified RNA-seq library purification. The amplified RNA-seq library was stored at -20°C for sequencing.

Sequencing data processing and quality control

All fractionated plasma samples isolated by SEC were randomized to reduce sample batch effects and were uniformly processed for RNA extraction, library preparation, and sequencing in Illumina flow cells. All 168 premade RNA-seq library sample were sequenced using NovaSeq 6000 system by Novogene company (Sacramento, CA). The premade RNA-seq library samples were equally distributed over three of NovaSeq S4 lanes for pair-end reads x 150 bp. Adapter sequences were trimmed using sickle tool [256]. After trimming, the quality of the reads were checked using FastQC (v0.11.7) [257, 258] and RSeQC (v2.6.4) [259]. Reads were aligned to the hg38.Ens_94.biomart human genome annotation using the STAR aligner (v2.5.3a) [115] with two pass mode flag. Duplicated reads were removed using the picard tool (v1.119) [260]. Read counts for each gene were calculated using the htseq-count tool (v0.11.2) [261] in intersection-strict mode. For each sample, we calculated exon, intron, intergenic fractions and protein coding fractions (CDS exons) using RSeQC [262].

5.4 Results

Characterization of Human Plasma Size Fractionation

In order to identify potential carriers of RNA in human plasma, we employed size-exclusion columns (SEC) to separate extracellular vesicles from non-vesicle associated fractions. Distinct physical properties of the individual fractions were characterized by tunable resistive pulse sensing technology (qNano), absorbance, dynamic light scattering (DLS), and transmission electron microscopy (TEM) (**Figure 5.1**). We measured concentration of extracellular vesicles (particles/ ml) along with the elution volume using

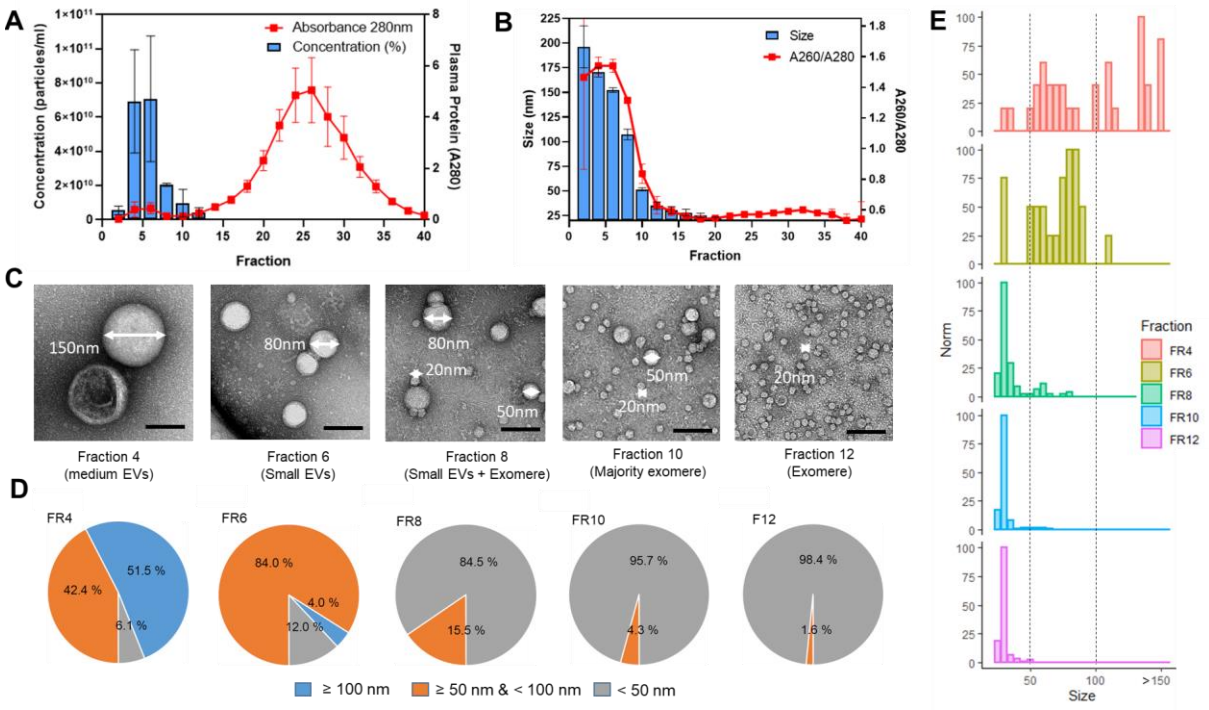


Figure 5.1 | Characterization of distinct EV subtypes through plasma fractionation
(A) Bar graph of EV concentration measured by tunable resistive pulse sensing using qNano on left-axis, and line plot of protein abundance measured by absorbance at 280 nm on right-axis with respect to plasma fractionation. X-axis indicates each fraction from size exclusion column. (B) Bar graph of mean hydrodynamic size (nm) distribution measured by dynamic light scattering on left-axis, and line plot of relative nucleic acid abundance measured by ratio of absorbance at 260 nm over 280 nm (A260/A80) on right-axis with respect to plasma fractionation. (C) Representative transmission electron microscopy images of particle subtypes in individual fraction containing medium EVs (fraction 4), small EVs (fraction 6), small EVs with exomeres (fraction 8), majority of exomeres (fraction 10), and exomeres (fraction 12) with scale bars, 100 nm. (D) Pie chart of percent distribution of particles with corresponding size ranges in fractions identified using TEM. (E) Histogram of particle size measured by transmission electron microscopy are shown for each fraction.

qNano (**Figure 5.1A**). Recovery of extracellular vesicles obtained from each fraction containing 2 ml volume resulted in range of 10^9 to 10^{10} vesicles per ml in fraction 2, 4, 6, 8, 10, and 12. Fraction 14 and onwards were below the detection limit of qNano. As expected, elution of vesicle peaks were found in fractions 4 and 6. We also analyzed elution of plasma soluble proteins using absorbance at 280 nm (A280) which had maximum peaks eluting in much later fractions 24 and 26 (**Figure 5.1A**). Additional measurement using dynamic light scattering (DLS) revealed hydrodynamic size of particles correlated with absorbance A260nm/A280nm ratio, indicating relatively high nucleic acid abundance in EV associated fractions (**Figure 5.1B**). In addition, size distribution measured by DLS indicated decreasing particle size with further elution fractions. To further characterize the distribution of heterogeneous EV sizes, we analyzed the morphology of EVs by TEM (**Figure 5.1C**). The proportion of distinct size ranges are shown for each fraction (**Figure 5.1D**). Fraction 4 contained the highest proportion (51.5%) of particle size ≥ 100 nm. Fraction 6 contained highest proportion (84%) of particle size between 50 nm and 100 nm. Fraction 12 contained highest proportion (98.4%) of particle size smaller than 50 nm (**Figure 5.1D**). Similar to qNano, fraction 14 and onwards composed of much higher amount of protein aggregates, which hindered direct visualization by TEM. The histogram of particle sizes displayed three distinct ranges: smaller than 50 nm, between 50 – 100 nm, and larger than 100 nm (**Figure 5.1E**). Based on qNano and TEM analysis, we further divided EV fractions into i) medium EVs (fractions 1-4: FR14) which major contribution from particles size ≥ 100 nm, ii) small EVs (fractions 5-8: FR58) which contained majority of particles between 50 and 100 nm, and iii) exomeres (fractions 9-12: FR912) which contained the majority of particles < 50 nm (**Figure 5.1E**). Utilizing A280 protein

adsorption, soluble protein components of plasma were divided into i) early protein peaks (fractions 16-19: FR1619), center of the protein peaks (fractions 23-26: FR2326), and iii) late protein peaks (fractions 30-33: FR3033) for downstream analysis (**Figure 5.1A**).

Transcriptomic Analysis of EVs and Non-vesicles

To determine the expression profile of cell-free RNAs across plasma fractions, we utilized aforementioned plasma fractions for our sample cohort which includes healthy donors (n=5), lung cancer (n=5), multiple myeloma (n=5), liver cancer (n=5), liver cirrhosis (n=4), and monoclonal gammopathy of undetermined significance (n=4). (**Figure 5.2A**). Across total of 168 fractionated plasma samples, total mean of 51.7 million (M) raw reads in the range of 25.9 M to 109.9 M were detected (**Supplementary Figure S5.1A, S5.1B**). After duplicate read removal, we observed a mean of 4.0 M uniquely mapped reads in the range of 70,152 to 34.2M. We found the percent of uniquely mapped reads declined towards protein-association fractions: 17.4 % (FR14), 10.5% (FR58), 4.0% (FR912), 5.0 % (FR1619), 4.4% (FR2326), and 1.6% (FR3033) (**Supplementary Figure S5.1C**). These findings were consistent with decrease of exon fraction maximum at 97% to minimum at 7% towards later protein fractions (**Supplementary Figure S5.2A**). While the global exon fraction declined towards protein enriched fractions, intron and intergenic fractions increased towards protein-enriched fractions (**Supplementary Figure S5.2B, S5.2C**). Among the ~12,000 distinct cfRNA transcripts found in fractionated plasma, 94 ± 0% of the cfRNA biotype were identified as protein coding transcripts which results were consistent across different clinical sample types and fractions (**Supplementary Figure S5.3A, S5.3B**).

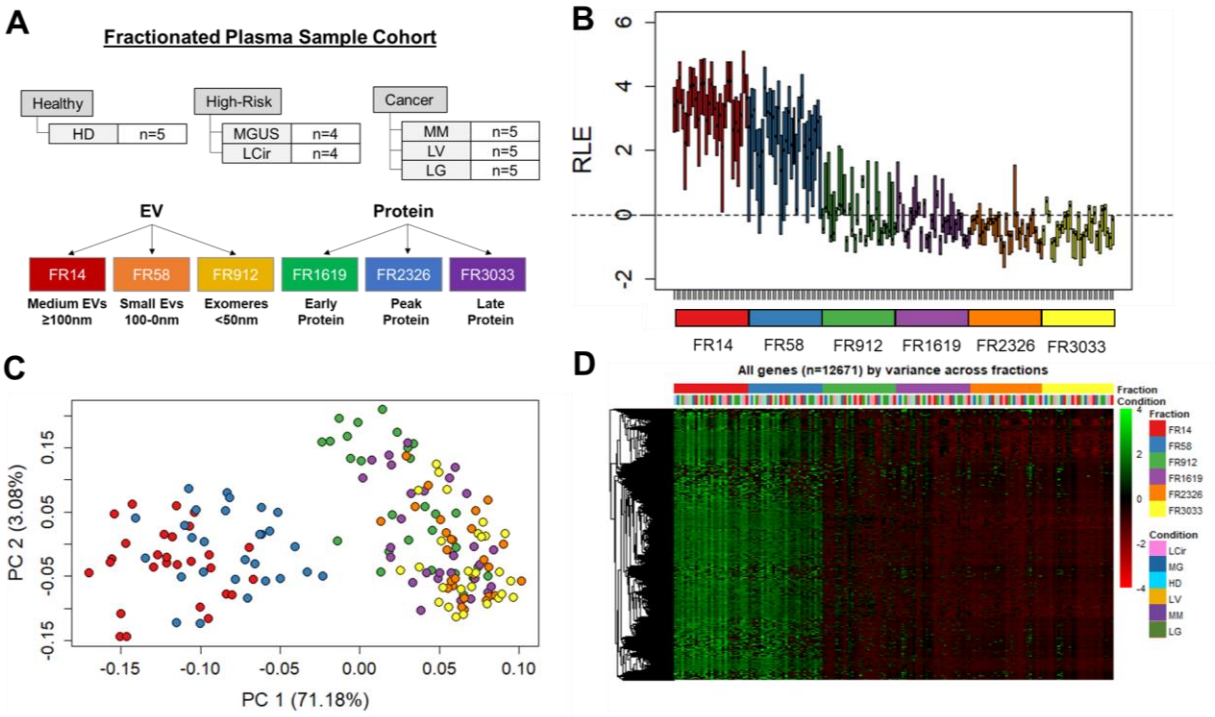


Figure 5.2 | Transcriptomic Analysis of EVs and Non-vesicles

(A) Schematic of total of 168 RNA sequencing samples derived from fractionated plasma. Plasma samples from healthy, high-risk (MGUS and LCir), and cancer patients (MM, LV, LG) were fractionated into EV and non-vesicle associated fractions. (B) Box plot of relative log expression of normalized read counts across EV and protein fractions. (C) Principal component analysis of top 500 genes with largest variance across individual fractions within healthy controls. (D) Heatmap expression of all genes from all conditions across all fractions revealing majority of cell-free mRNA are within FR14 and FR58.

In order to identify gene expression differences across all fractions within patient cohorts, we plotted boxplot of relative log expression across 168 fractionated plasma samples (**Figure 5.2A, 5.2B**). The results revealed the majority of cell-free mRNAs is found in FR14 and FR58 corresponding to medium and small EVs in contrast to exomere and protein enriched fractions (**Figure 5.2B**). Unsupervised principal component analysis

(PCA) was conducted using top 500 genes with the largest variance. The first two principal components of PCA clearly separated medium and small EVs from other plasma fractions (**Figure 5.2C**). Based on PCA, the exomere enriched fraction (FR912) exhibited a higher degree of similarity to protein enriched fractions than medium or small EV enriched fractions. Agreeing with this finding, a heatmap analysis performed on all detected genes displayed that cf-mRNA were predominantly enriched in medium EV and small EV enriched fractions (**Figure 5.2D**). Similar results were found when hierarchical clustering of expression profiles was plotted per disease type (**Supplementary Figure S5.4A-S5.4F**).

Cell-free mRNA are Present and Protected in Extracellular Vesicles

To evaluate if other potential carriers of RNA such as lipoproteins, and RNA binding proteins cofractionate with EV fractions, we performed immunoprecipitation using antibodies against canonical EVs markers (CD9), apolipoprotein (APOA1) and Argonaut complexes (Ago2) (**Figure 5.3A**). CD9 was preferentially enriched in FR14 and FR58 confirmed the presence of EVs. In contrast, APOA1 and Ago2 were enriched in protein fractions and showed no measureable level in EV fractions (FR14 and FR58). To examine whether EVs protect cf-mRNAs from endogeneous RNase in human plasma, we treated the control RNA (total human liver tissue RNA) and EV fraction with either RNase A, detergent Triton X-100 to disrupt membrane bound EVs, and both RNase A and Triton X-100 or buffer alone. Negative raw ct value was plotted against different treatment for control RNA and EVs (**Figure 5.3B**). When treated with RNase A alone, we found control RNA is degraded similar to when RNase A and Triton X-100 were treated together. However, PCR signal for vesicles treated with RNase A remains similar to when buffer or

triton X-100 was treated alone. In addition, vesicle derived RNA treated with both RNase A and Triton X-100 led to near total digestion of RNAs, confirming that circulating cf-mRNAs are protected in membrane bound vesicles. Taken together, our results indicate that cell-free mRNAs are present and protected within extracellular vesicles in plasma.

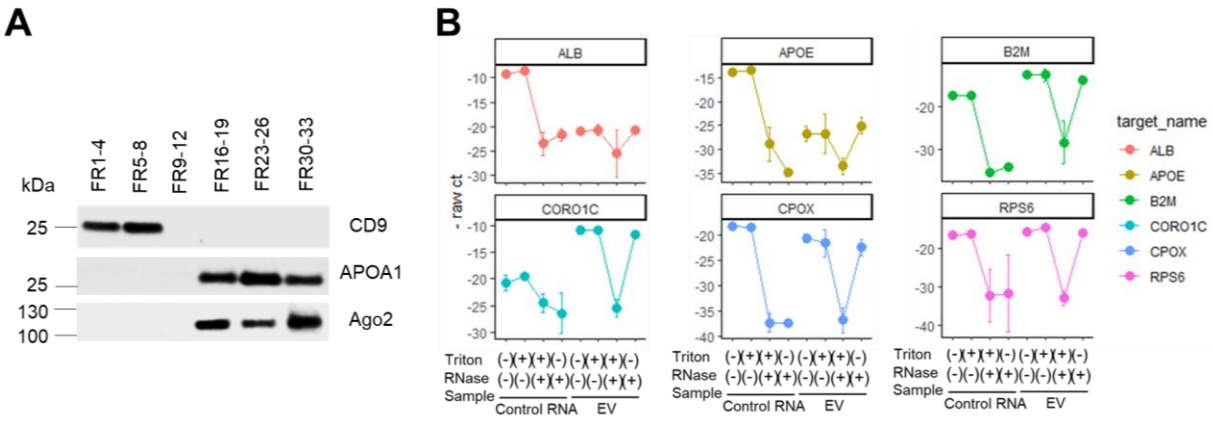


Figure 5.3 | Relative quantification of RNA by qRT-PCR and immunoprecipitation (A) Expression of protein markers (CD9, APOA1, and Ago2) using immunoprecipitation across plasma fractions. (B) A line plot of negative raw ct of individual gene with RNase and/or detergent using qRT-PCR. RNA isolated from EV fraction and control RNA from three healthy individuals were treated with RNase and/or detergent.

Selective Packaging of Cancer Differentiating Genes

We hypothesized that cancer may dysregulate the mRNA content of circulating EV subpopulations. To identify the enrichment of cancer differentiating genes in specific fractions, hierarchical clustering analysis was performed on significantly differentially expressed genes between healthy and cancer per plasma fraction. We found specific genes significantly upregulated in lung cancer that are contained within medium EVs, small EVs, and exomeres (**Figure 5.4A**). In addition, fewer sets of lung cancer differentially expressed genes were identified in early, peak, and late eluting protein fractions (**Supplementary**

Figure S5.5A). In order to predict the classification, linear discriminant analysis (LDA) was performed on lung cancer differentially expressed (DE) gene sets per fraction. By employing leave-one-out cross validation (LOOCV), we revealed individual samples trained on the DE gene sets identified from EVs were more accurately classified compared to DE gene sets identified from protein enriched fractions (**Figure 5.4B, Supplementary Figure S5.5B**). Moreover, we also identified genes present in EV and protein fractions differentially expressed for liver cancer and for multiple myeloma (**Supplementary Figure S5.6A and S5.7A**). To assess if cancer distinguishing genes exhibited global or selective increase across individual fractions, we characterized the number of genes which were unique or shared across individual fractions. This revealed that majority of differentially expressed genes were found in unique fractions (**Supplemental Figure S5.8**). In order to compare the relative gene expression profiles in healthy and cancer plasma across fractions, we performed hierarchical clustering analysis on the identified DE gene sets. Unique cancer distinguishing genes per fraction were assigned as clusters corresponding to their fraction enrichment (FR14, FR58, FR912, FR1619, FR2326, FR3033 as clusters 1-6 respectively). Intriguingly, our supervised clustering analysis revealed six distinct groups whose gene expressions were enriched in cancer relative to healthy in a specific fraction (**Figure 5.4C**). Majority of the cancer differentiating genes were found in FR14 and FR58 enriched in medium and small EVs. We found a similar selective enrichment of cancer differentiating genes identified in distinct EV or protein fractions for multiple myeloma and liver cancer (**Supplementary Figure S5.6B and S5.7B**). Our results revealed cancer-associated cell-free RNA signatures are distinctively packaged in patient plasma.

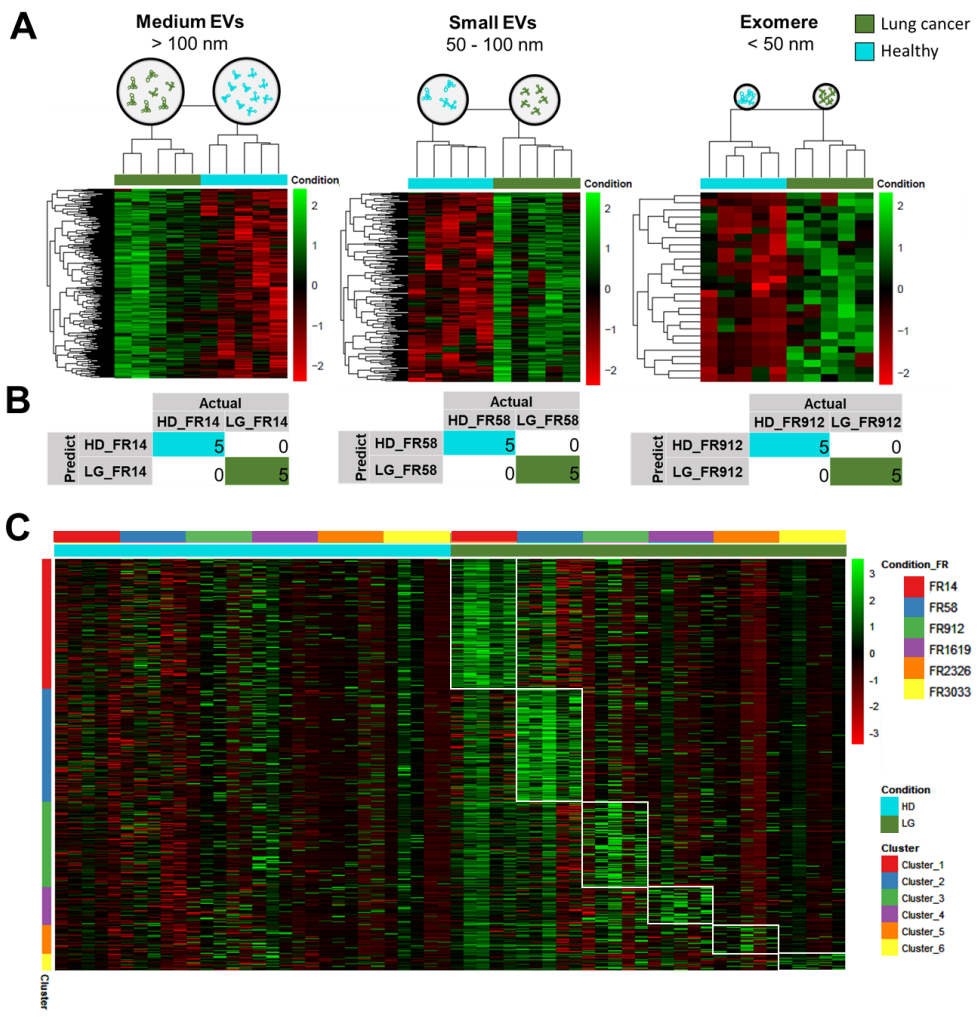


Figure 5.4 | Distinct cancer differentiating cell-free mRNA across fractions in human plasma

(A) Heatmap of log counts of lung cancer differentially expressed genes within individual fractions enriched in medium EVs, small EVs, and exomeres were compared between lung cancer and healthy. Differentially expressed genes which showed statistical significance (student's t test, p-value < 0.05) were used. (B) Leave-one-out cross validation testing accuracy of linear discriminant analysis algorithm for classification using lung cancer upregulated genes specific to individual fraction. (C) Heatmap of gene expression in lung cancer relative to healthy across fractions. A total of 800 significantly differentially expressed genes were used. Clusters were assigned to genes corresponding to their enriched fraction based on log 2 fold changes (FR14, FR58, FR912, FR1619, FR2326, FR3033 as clusters 1-6 respectively).

To assess the potential roles of these unique cancer distinguishing gene sets revealed per fraction, we performed gene set enrichment analysis (GSEA) curated on biological, chemical and genetic perturbation. We found that genes enriched in FR14 (medium EVs) derived from liver cancer patients shared genes matching the human liver cell atlas [298] (**Supplementary Figure S5.9A**). Additionally, our liver cancer differentiating genes in medium EVs (i.e. C1S, F5, and ADH1B) were found in human liver cell atlas and liver tissue specific expression analysis [299]. Other gene sets including MR1, RGS5, XXXXYLT1, FAT1, CYSTM1, CNEP1R1, VAMP5, MYNN, CPVL, and PALLD identified in medium EVs (FR14) of liver cancer plasma were upregulated in hepatocellular carcinoma patient tissues compared to normal liver samples [300]. The differentiating gene cluster associated with small EVs (FR58) also contained liver-specific genes related to hepatocyte differentiation [298, 299, 301, 302] (**Supplementary Figure S5.9B**). Notably, lung cancer differentiating genes in medium EVs (i.e. KIF2C, PSAT1, CCNA2, SCD, DTYMK, PFN2, and CDCA8) were associated with lung cancer poor survival prognosis [303]. (**Supplementary Figure S5.9C**). Finally, for multiple myeloma, genes upregulated in the medium EV cluster (FR14) were associated with epithelial mesenchymal transitions, a hallmark of increased aggressiveness, invasion, and metastatic potential [304] (**Supplementary Figure S5.9D**). Overall, our gene sets uniquely enriched in each fraction revealed relevant biological significance associated with aberrant gene sets identified in corresponding tumor tissue samples and predicted cancer patients' poor survival outcome.

Specific cfRNA Signatures Associated with High-risk Group and Cancer

Next, to assess the potential of distinct cf-mRNA carriers in cancer progression, we investigated the selective packaging of transcriptomic signatures associated with high-risk groups (liver cirrhosis and MGUS) and their corresponding cancer types (liver cancer or multiple myeloma, respectively). We performed pairwise comparisons between healthy, high-risk groups, and their corresponding cancer types by associated fraction. By combining these DE gene sets identified from each fraction, we found that there are 6 patterns of differentiating genes, denoted as clusters: clusters 1-3 included genes uniquely enriched for specific conditions (healthy, high-risk group, or cancer respectively), and clusters 4-6 which included genes enriched in paired conditions (i.e. high-risk and cancer in cluster 6) (**Figure 5.5A, 5.5D**). In order to reflect genes associated with each fraction from these clusters, the number of differentially expressed genes in each fraction per cluster was generated (**Figure 5.5B, 5.5E**). Interestingly, healthy-upregulated genes (cluster 1) were mostly found in the protein enriched fraction FR2326. In contrast, high-risk and cancer upregulated genes (cluster 2 and 3 respectively) were mostly found in medium and small EVs (FR14 and FR58). Specifically, liver cancer upregulated genes were preferentially found in small EVs (FR58), while MM upregulated genes were enriched in medium EVs (FR14) (**Figure 5.5B, 5.5E**). Although further studies including a larger sample size are required to further validate this model, our findings constitute a proof-of-principle that cell-free mRNA in human plasma are selectively found in extracellular vesicles and that their packaging differences are associated with cancer progression.

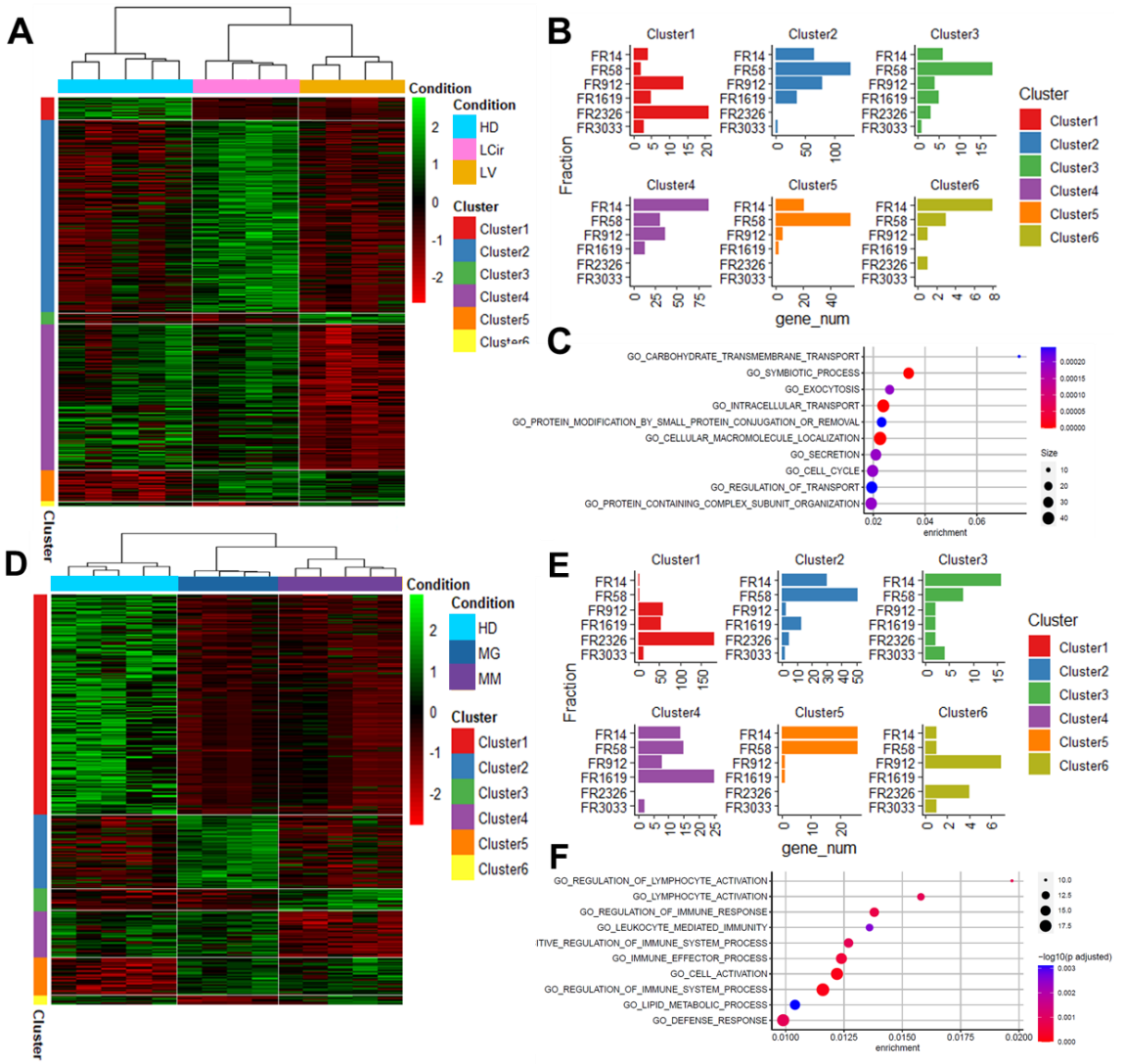


Figure 5.5 | Specific cfRNA signatures associated with high risk group and cancer

Heatmap of log counts of differentially expressed genes (A) between healthy (HD), liver cirrhosis (LCir) and liver cancer (LV) and (D) between HD, monoclonal gammopathy of undetermined significance (MGUS) and multiple myeloma (MM). Differentially expressed genes (student's t-test, p-value < 0.01) within individual fraction were identified by pair-wise comparison which results in 6 distinct patterns. Representative bar plot of number of genes identified in each cluster across fraction (B) between HD, LCir, and LV and (E) between HD, MGUS, and MM. Gene set enrichment analysis (GSEA) was performed on (C) LCir upregulated genes and (F) MGUS upregulated genes. GSEA was performed from Molecular signatures database using C5:BP derived from biological process ontology.

To relate gene products in terms of biological properties, we performed gene ontology on 6 clusters identified for high-risk groups. In the case of healthy, liver cirrhosis and liver cancer comparisons, cluster 1 contained gene sets regulating organelle organization (**Supplementary Figure S5.10A**). Cluster 2 which is specific to liver cirrhosis, was found to be involved with exocytosis, secretion, and regulation of transmembrane transports (**Figure 5.5C**). Cluster 3, which is specific to liver cancer, was found to be involved with chylomicron remodeling and lipoprotein particle remodeling which is a major function of liver (**Supplementary Figure S5.10B**). Cluster 4, which are upregulated in both healthy and liver cirrhosis, contained regulation of response to stimulus and signaling (**Supplementary Figure S5.10C**). Cluster 5, which is both upregulated in liver cirrhosis and liver cancer, contained gene sets involved in the regulation of fatty acid transport and regulation of cell death (**Supplementary Figure S5.10D**). In the case of healthy, MGUS and multiple myeloma comparisons, cluster 1 was involved in regulation of cellular transport and organization (**Supplementary Figure S5.11A**). Cluster 2, which is specific to MGUS, was found to be involved with immune system processes and lymphocyte activation (**Figure 5.5F**). Cluster 3, which is specific to multiple myeloma, was found to be involved with oxygen transport and blood coagulation (**Supplementary Figure S5.11B**). Cluster 4, which are upregulated in both healthy and MGUS, was involved in protein targeting to membrane or endoplasmic reticulum (**Supplementary Figure S5.11C**). Lastly, cluster 5 which is both upregulated in MGUS and multiple myeloma contained gene sets involved in metabolic processes (**Supplementary Figure S5.11D**).

5.5 Discussion

Although recent studies have supported circulating cf-mRNA as promising cancer-differentiating biomarkers, how these RNAs reside within plasma remains unknown. Discerning which types of RNA are being carried in the context of disease progression will be highly valuable for disease diagnosis. While numerous reports have shown miRNA cargo types, studies on circulating mRNA cargo types is limited. Recent studies have revealed different potential extracellular miRNA cargo types, including both EVs and non-vesicle carriers [39, 138]. The NIH extracellular RNA communication consortium created exRNA atlas resource which major carrier of miRNA were extensively compared across 19 studies [138]. Comparative statistical studies on miRNA carrier subclasses were determined, revealing distinct miRNA biotype composition within each cargoes [79]. By investigating potential cell-free mRNA carrier into each category (medium EVs, small EVs, exomeres, and early-, middle-, and late-eluted protein fractions) using size exclusion column, we discovered that 98.9% of mRNAs in the circulation are present in vesicle associated fraction.

To our knowledge, this is the first study analyzing cf-mRNA contents within fractionated plasma, allowing for characterization into their respective extracellular vesicle or soluble plasma protein fractions. Although total of 168 isolated RNA from fractionated samples revealed uniform processing and input reads, we observed the significant difference in relative log expression between extracellular vesicles and soluble plasma proteins. Synthetic spike-in RNA is utilized in many studies as processing and normalization controls [295, 305]. Normalization using synthetic RNA spike-in control enabled accurate assessment of cell-free mRNA across fractions, revealing EVs are

primary carrier of cell-free mRNAs. Further characterization using RNase treatment with and without detergent to disrupt the membrane bound form revealed remarkable stability of circulating mRNA is attributed to extracellular vesicles. To assess different RNA carriers, we used immunoprecipitation on well-characterized non-vesicular carriers such as lipoproteins (APOA1) and the RNA binding protein (Ago2) as being enriched in protein fractions, revealing those were not the major types of circulating mRNA carriers.

Previous reports have shown miRNA found in human plasma were primarily found in ribonucleoprotein complexes, revealing Argonaut2 complexes as major carrier of miRNAs [39]. Arroyo et al. confirmed Argonaut2 was observed in the plasma fraction coeluted with other miRNAs (miR-16, miR-92, and miR-122) [39]. However, some miRNA, let-7a, which might originate from cell types known to generate vesicles could also detected in EVs [39]. Other studies revealed almost all of the miRNAs in normal human plasma could be immunoprecipitated by Ago2 antibodies [138, 292]. In contrast, another study following similar immunoprecipitation protocol revealed presence of Ago2 in EV and detected miRNA associated with EVs [286]. We further investigated this controversial result and revealed Ago2 is mainly detected in protein enriched fraction in the absence of lysis buffer meanwhile Ago2 can also be detected inside EV when EVs were lysed. Treatment of isolated EV samples with or without lysis buffer critically affected the abundance of Ago2 in human plasma. Therefore, associated study of exRNA carriers should be carried by taking this effect into consideration.

It is not well known whether heterogeneous EV cargo composition significantly vary due to the disease progression. While numerous reports have shown functional delivery of miRNA by extracellular vesicles promotes tumorigenesis, invasion, and cell

proliferation in culture medium [306, 307], studies validating using clinical samples is limited. Importantly, the presence of oncogenes such KRAS has been shown to suppress Ago2 interactions with endosomes, resulting in different miRNA secretion into exosomes [63]. Selective sorting of different miRNA into vesicles in cancer cell lines supports our hypothesis that selective packaging of cancer differentiating genes can be found in EVs. Interestingly, we found majority of cancer upregulated genes were packaged within medium EV and small EV enriched fraction while cancer downregulated genes were packaged in protein-enriched fraction. Additional analysis was performed to investigate whether high-risk group can be distinguished from healthy and cancer. Our results highlighted 6 distinct patterns that may be altered due to disease progression. Collectively, our results suggest how ensembles of genes are either shared or distinctively dysregulated throughout disease progression.

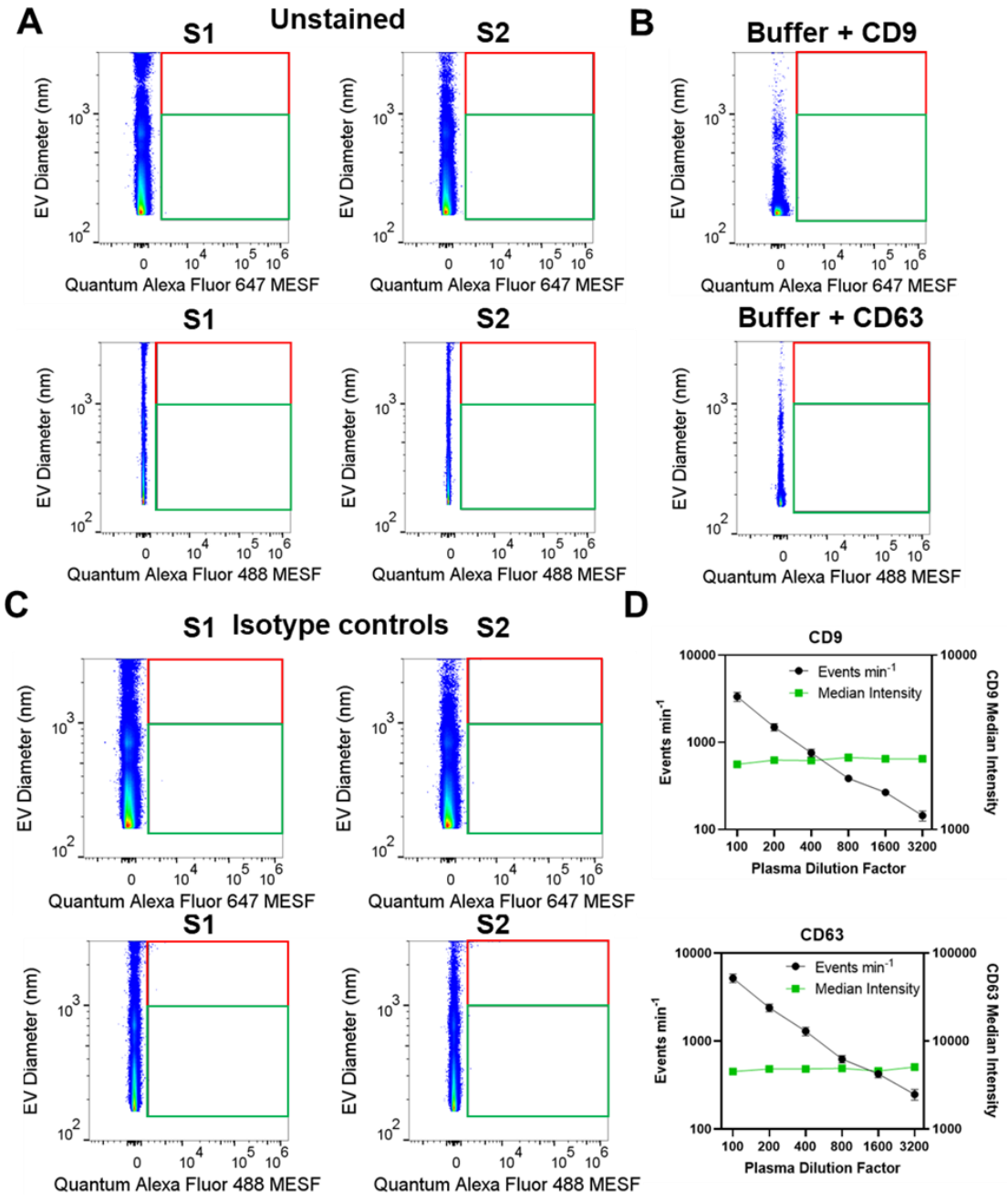
Although larger clinical cohorts are necessary to investigate the clinical potential of this approach, gene set enrichment analysis revealed the identified gene sets enriched in each cluster were also overlapped with relevant clinical studies with larger cohorts [298-303]. From chemical and genetic alteration, we found our gene sets enriched in EV fractions showed an overlap with RNA upregulated tumor tissue samples for non-small cell lung cancer and hepatocellular carcinoma [298-303]. Gene sets identified in EV fraction from multiple myeloma was associated with epithelial mesenchymal transition (EMT) associated with increased aggressiveness, invasion, and metastatic potential [304]. Overall, our gene sets identified in unique fractions revealed relevant biological significance potentially associated with aberrant tissue specific genes across cancer.

5.6 Conclusions

In conclusion, comprehensive investigation of cell-free mRNA distribution across size-based fractionated plasma was performed from three different cancer types (lung cancer, liver cancer, and multiple myeloma) as well as high-risk group (liver cirrhosis and MGUS), highlighting the important roles of EVs as potential cell-free mRNA carriers. These results presented here will serve as valuable resource in understanding the remarkable stability of circulating mRNA attributed to extracellular vesicles as well as how dysregulated RNA packaging into different sizes of EVs were found in human plasma.

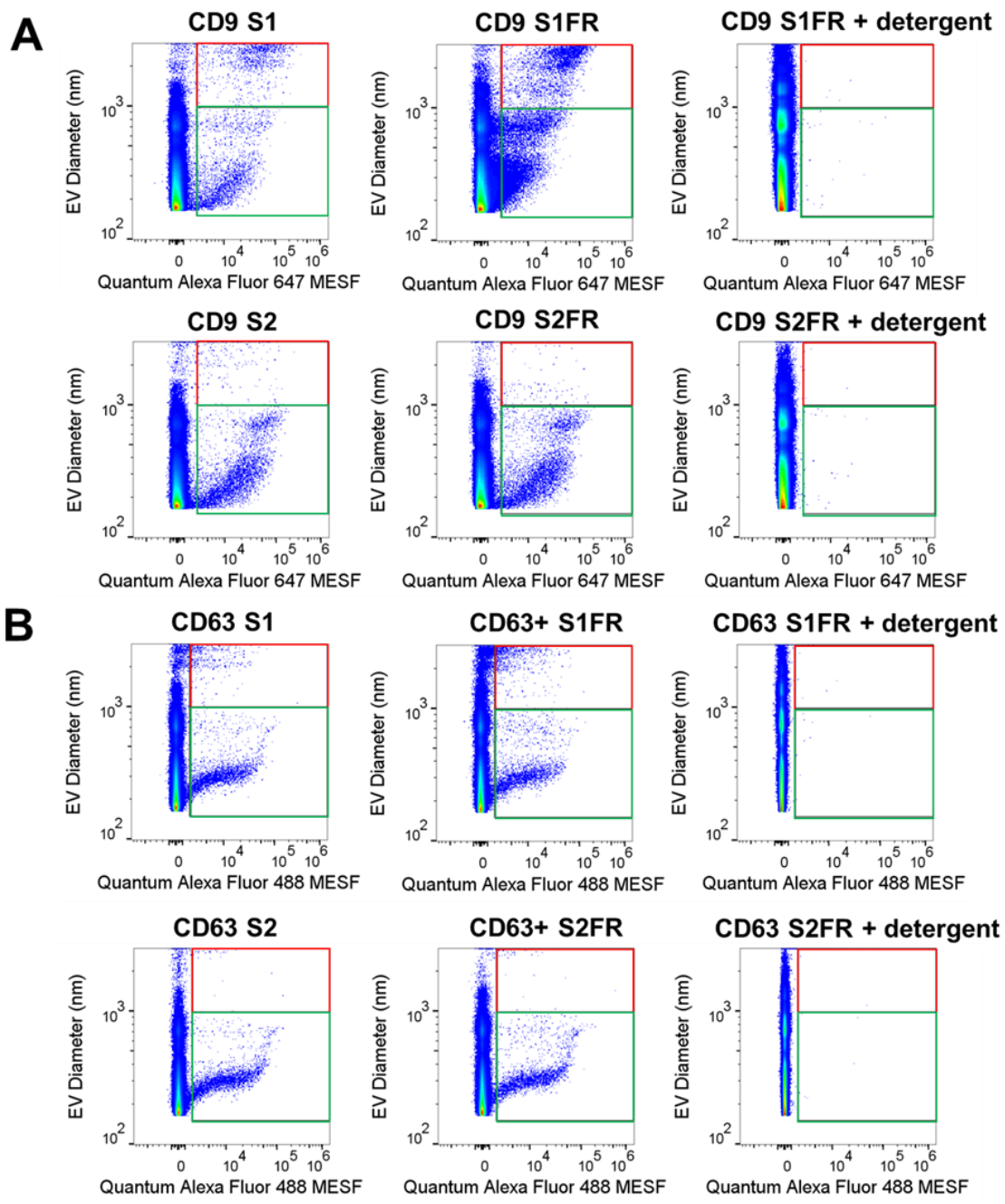
Chapter five is a draft of the manuscript entitled: “Selective packaging of extracellular vesicles RNA association with cancer progression”, Hyun Ji Kim, Breeshey Roskams-Hieter, Matthew Rames, Josephine Briand, Josiah Wagner, Aaron Doe, and Thuy T. M. Ngo, In preparation (2021). The author of this dissertation is the first author of this manuscript.

Appendix A: Supplementary Tables and Figures



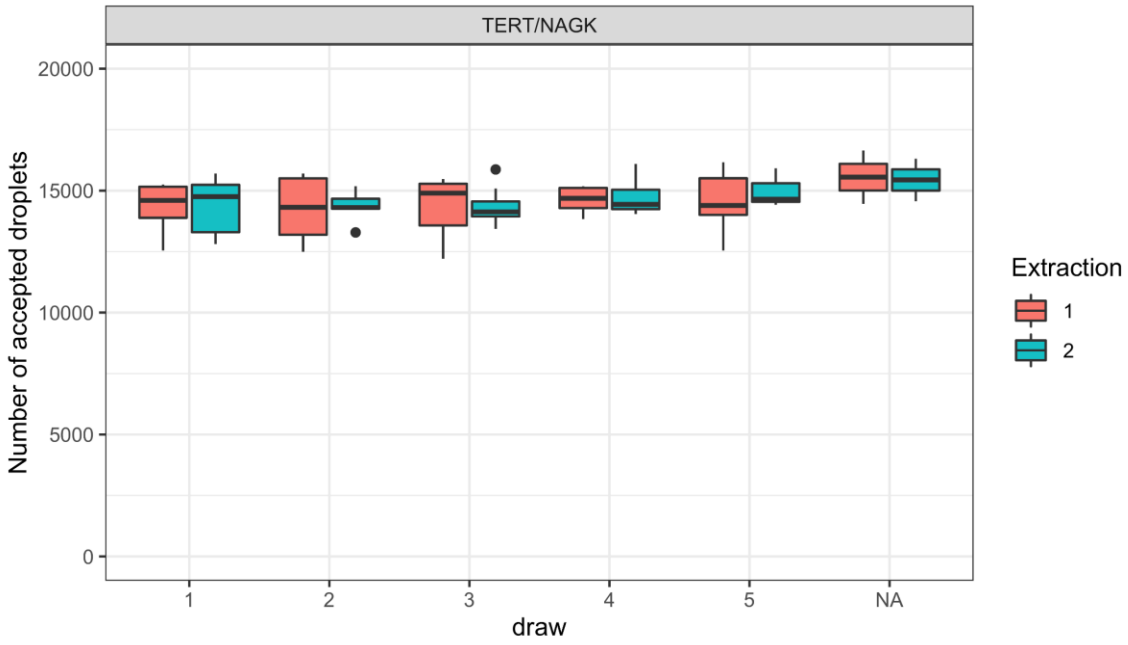
Supplementary Figure S2.1 | Flow cytometry experimental assay controls

Representative flow cytometry dot plot of antibody with (A) unstained S1 and S2 plasma, (B) Alexa Fluor 647 conjugated CD9 or Alexa Fluor 488 conjugated CD63 with buffer alone, and (C) Alexa Fluor 647 conjugated or Alexa Fluor 488 conjugated isotypes in S1 and S2 plasma. (D) Scatter plots of $CD9^+$ and $CD63^+$ EVs dilution controls.



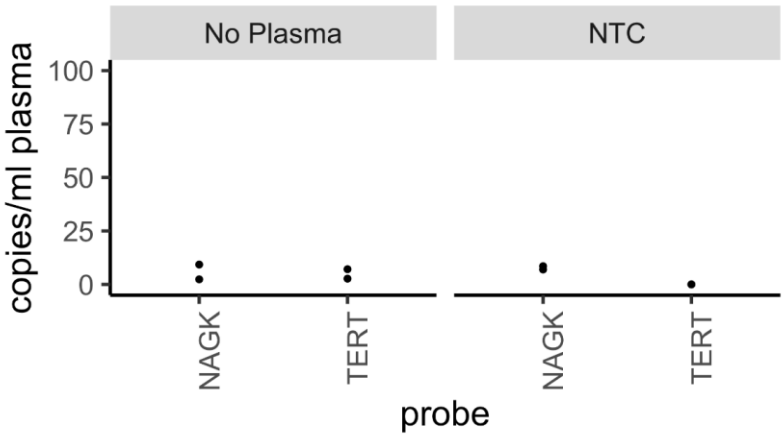
Supplementary Figure S2.2 | Freeze thaw effect and plasma EV detergent treatment

Representative dot plots EVs from S1 and S2 plasma, respective freeze-thaw processing (S1FR and S2FR), and detergent controls on (A) $CD9^+$ EVs and (B) $CD63^+$ EVs. S1FR and S2FR were treated with detergent (2% SDS) prior to staining.

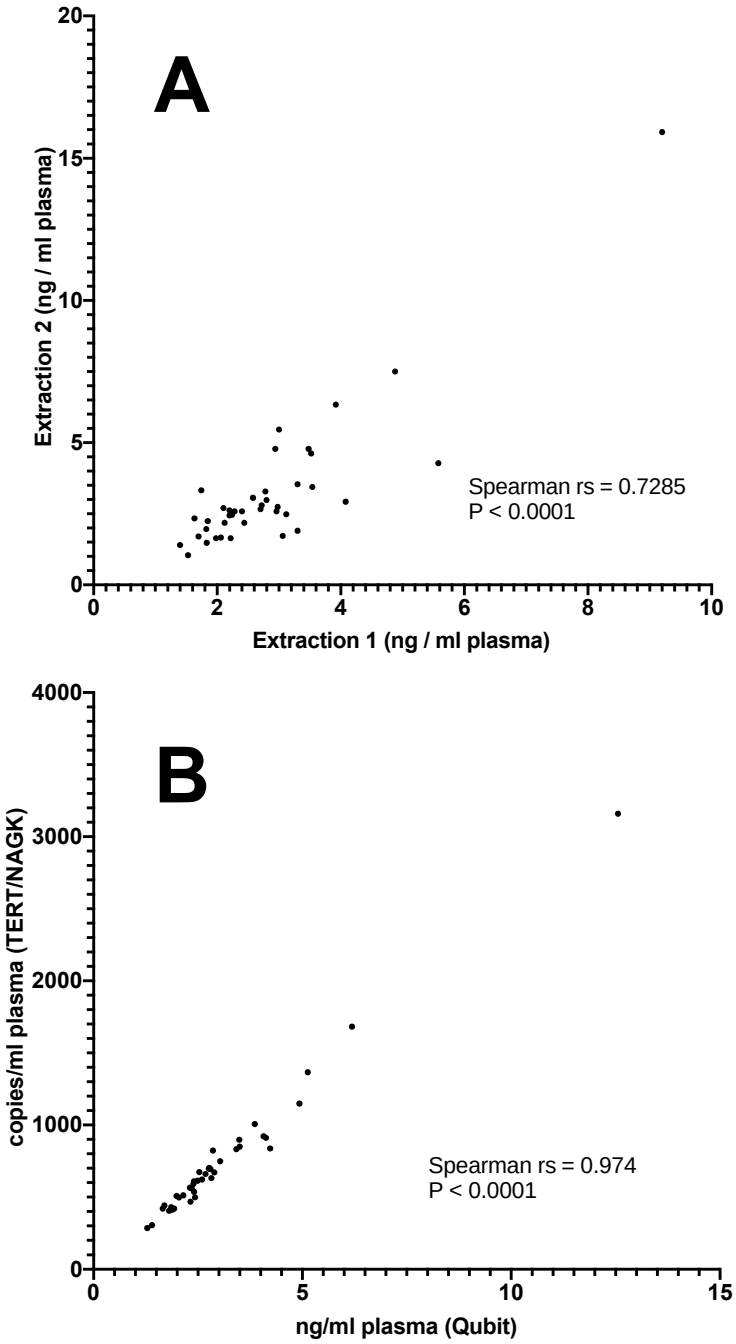


Supplementary Figure S3.1 | Total number of droplets accepted by the QX200 ddPCR droplet reader for nucleic acid quantitation.

Number of droplets for each draw are shown for cfDNA ddPCR analysis. Technical replicates for each sample were averaged.

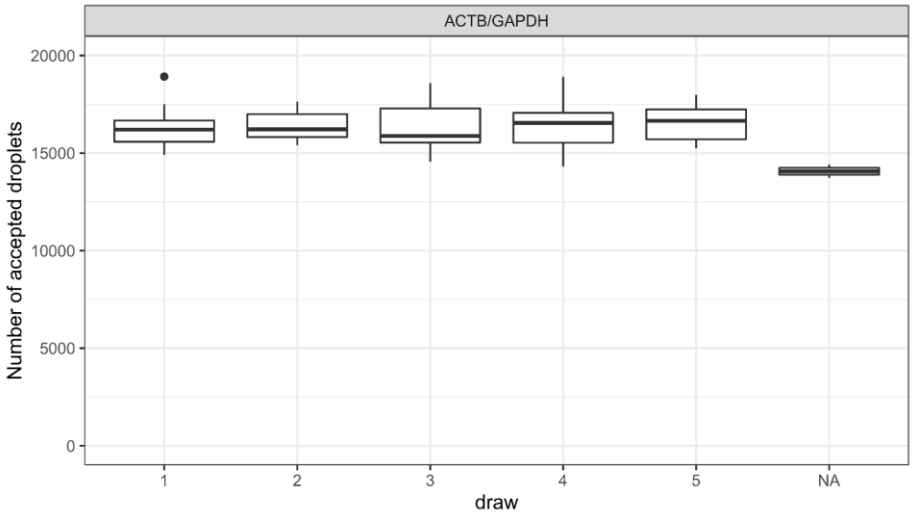


Supplementary Figure S3.2 | Negative controls for ddPCR measurement of cfDNA. Data points are from the two independent cfDNA extractions performed in this work. Negative controls were measured using ddPCR at the same time as the plasma samples



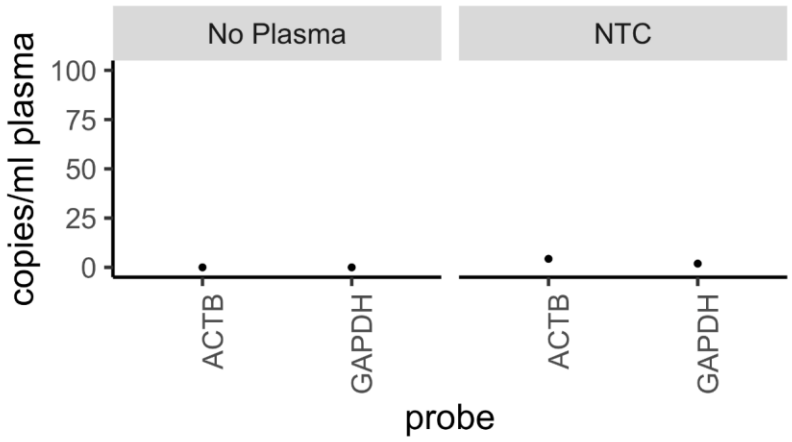
Supplementary Figure S3.3 | Nonparametric spearman correlations

Nonparametric Spearman correlation coefficients (rs) calculated between cfDNA extraction 1 and extraction 2 using Qubit measurements (A) and between Qubit and ddPCR (TERT and NAGK averaged) measurements of cfDNA (B). The correlations for both comparisons were statistically significant.



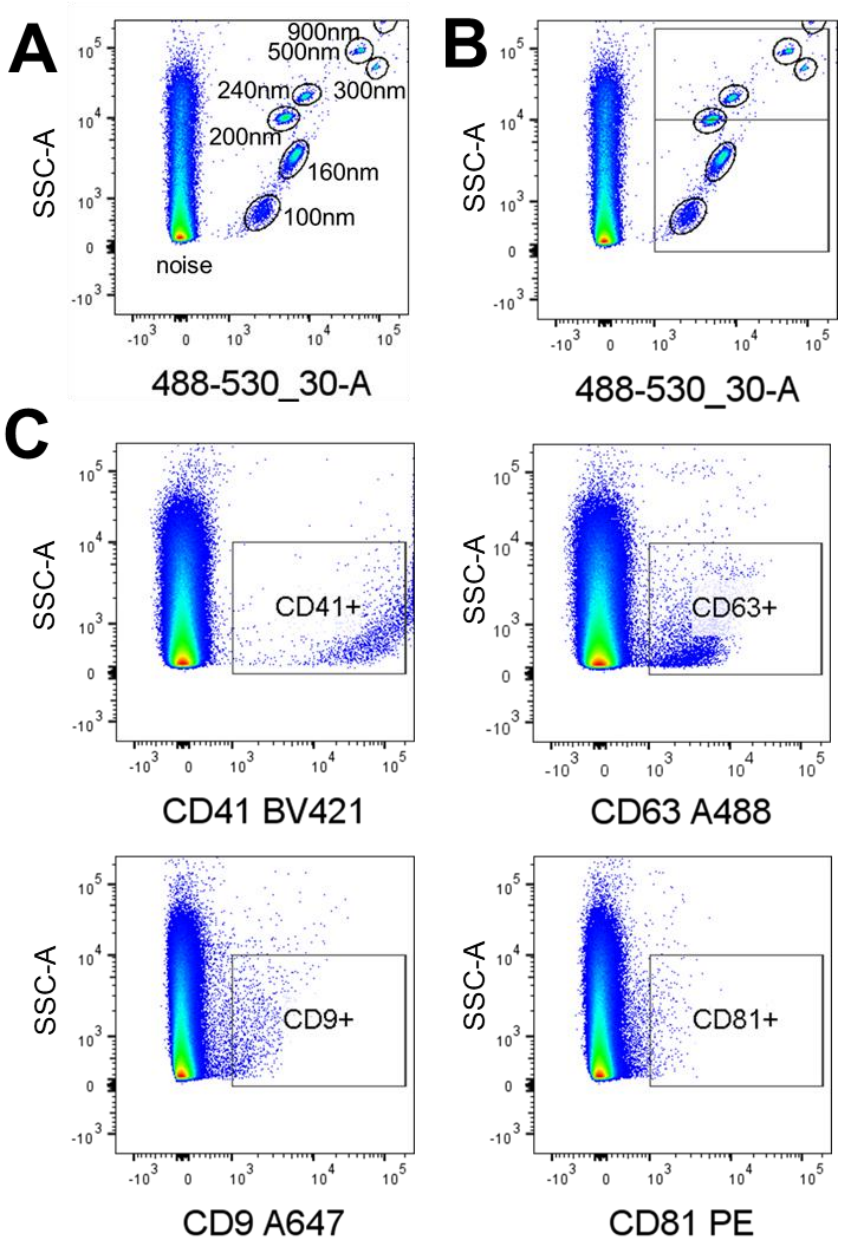
Supplementary Figure S3.4 | Total number of droplets accepted by the QX200 ddPCR droplet reader for nucleic acid quantitation.

Number of droplets for each draw are shown for cfRNA ddPCR analysis. Technical replicates for each sample were averaged.



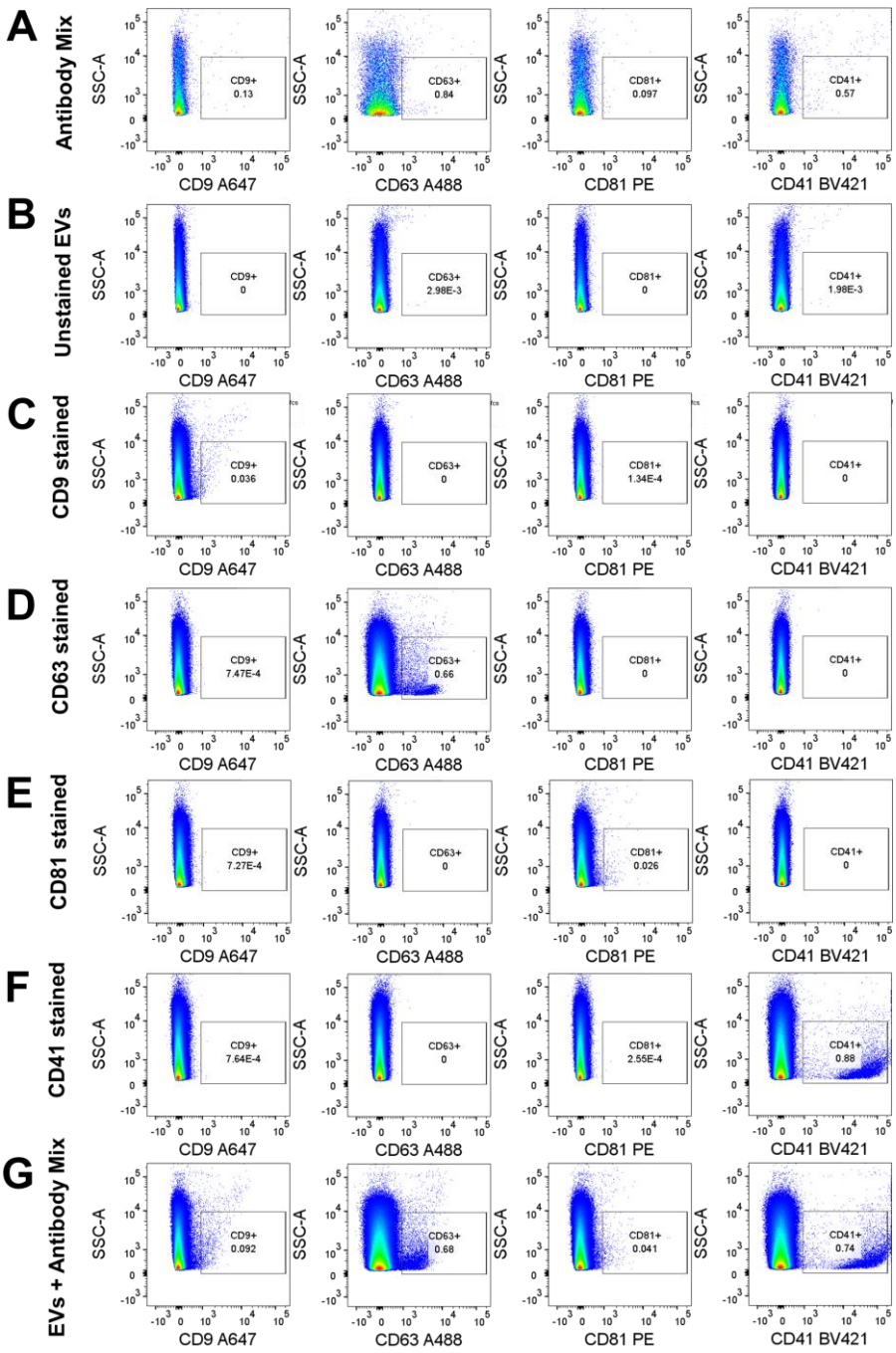
Supplementary Figure S3.5 | Negative controls for ddPCR measurement of cDNA derived from cfRNA.

Negative controls were measured using ddPCR at the same time as the plasma samples.



Supplementary Figure S3.6 | Flow cytometry set-up and gating for detection of plasma EVs.

(A) Calibration of SSC and fluorescence signal using polystyrene beads as a relative size reference shown in SSC/FL plot. Side scatter (SSC) intensity was set at 10^4 aligned to 200-nm bead reference. (B) Positioning of calibration beads with respect to gating areas for plasma EVs in SSC/FL plot. (C) Representative plasma EVs detected by CD41, CD63, CD9, and CD81 fluorescence for one blood draw of patient HD4.



Supplementary Figure S3.7 | Flow cytometry assay controls for plasma EV measurement.

Dot plot of SSC/FL with respective assay controls: (A) antibody mix only, (B) unstained plasma EVs, (C) CD9 stained plasma EVs, (D) CD63 stained plasma EVs, (E) CD81 stained plasma EVs, (F) CD41 stained plasma EVs, and (G) plasma EVs stained with antibody mix.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	1.28	2.39	2.32	1.81	1.82
Median[a]	1.88	3.46	3.56	2.47	2.89
Maximum[a]	2.49	4.13	12.56	2.86	4.61
Mean[a]	1.82	3.31	4.61	2.47	3.05
Std. Deviation[a]	0.34	0.64	3.10	0.33	1.20
Lower 95% CI of mean[a]	1.58	2.85	2.39	2.23	1.14
Upper 95% CI of mean[a]	2.07	3.76	6.82	2.71	4.96

Post-hoc pairwise comparison	P-value	Adjusted P	Summary[b]
HD1 - HD2	0.005	0.030	*
HD1 - HD3	0.020	0.030	*
HD1 - HD4	0.016	0.030	*
HD2 - HD3	0.142	0.142	ns
HD2 - HD4	0.016	0.030	*
HD3 - HD4	0.040	0.049	*

[a]ng per ml plasma

[b] *, $P < 0.05$; ns, not significant.

Supplementary Table S3.1 | Total plasma cfDNA concentration summaries and statistics as measured by Qubit.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	329.5	541.8	476.9	420.5	461.2
Median[a]	452.2	878.9	779.8	612.5	722
Maximum[a]	610.3	1018	3282	846.4	1129
Mean[a]	461.2	815.5	1129	628.5	758.6
Std. Deviation[a]	90.9	150.3	870.2	139.2	286.2
Lower 95% CI of mean[a]	396.2	708	506.5	528.9	303.1
Upper 95% CI of mean[a]	526.3	923	1752	728.1	1214

Post-hoc pairwise comparison	P-value	Adjusted P	Summary[b]
HD1 - HD2	0.006	0.035	*
HD1 - HD3	0.027	0.048	*
HD1 - HD4	0.032	0.048	*
HD2 - HD3	0.176	0.176	ns
HD2 - HD4	0.032	0.048	*
HD3 - HD4	0.060	0.073	ns

[a]copies per ml plasma

[b] *, $P < 0.05$; ns, not significant.

Supplementary Table S3.2 | Plasma *TERT* concentration summaries and statistics as measured by ddPCR.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	242.1	569.7	430.3	389.7	407
Median[a]	409.1	786.6	779.5	603.3	692
Maximum[a]	619.4	997.1	3037	800.5	1103
Mean[a]	407	798.3	1103	585.6	723.5
Std. Deviation[a]	106.6	141.4	771.9	106.2	299.3
Lower 95% CI of mean[a]	330.7	697.1	551	509.6	247.2
Upper 95% CI of mean[a]	483.2	899.5	1655	661.6	1200

Post-hoc pairwise comparison	P-value	Adjusted P	Summary[b]
HD1 - HD2	0.005	0.029	*
HD1 - HD3	0.023	0.035	*
HD1 - HD4	0.019	0.035	*
HD2 - HD3	0.177	0.177	ns
HD2 - HD4	0.015	0.035	*
HD3 - HD4	0.051	0.061	ns

[a]copies per ml plasma

[b] *, $P < 0.05$; ns, not significant.

Supplementary Table S3.3 | Plasma *NAGK* concentration summaries and statistics as measured by ddPCR.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	1.036	0.966	0.805	0.777	1.439
Median[a]	1.428	1.638	1.369	1.663	1.558
Maximum[a]	2.576	2.072	2.667	2.072	1.575
Mean[a]	1.57	1.546	1.439	1.575	1.533
Std. Deviation[a]	0.492	0.3864	0.511	0.406	0.0636
Lower 95% CI of mean[a]	1.218	1.27	1.073	1.284	1.431
Upper 95% CI of mean[a]	1.922	1.823	1.804	1.866	1.634

[a]ng per ml plasma

Supplementary Table S3.4 | Total plasma cfRNA concentration summaries and statistics as measured by Bioanalyzer.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	4,739	7,230	12,188	7,812	18,632
Median[a]	19,413	33,055	20,199	27,020	25,652
Maximum[a]	28,658	44,420	39,692	64,888	30,153
Mean[a]	18,632	29,989	21,314	30,153	25,022
Std. Deviation[a]	7,069	12,483	9,323	15,813	5,932
Lower 95% CI of mean[a]	13,575	21,059	14,645	18,841	15,582
Upper 95% CI of mean[a]	23,689	38,918	27,983	41,466	34,462

[a]copies per ml plasma

Supplementary Table S3.5 | Plasma *ACTB* cDNA concentration summaries and statistics as measured by ddPCR.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	786	1,498	3,040	1,284	4,532
Median[a]	5,821	9,666	4,353	5,479	5,478
Maximum[a]	7,656	11,525	7,582	8,839	8,444
Mean[a]	5,597	8,444	4,532	5,359	5,983
Std. Deviation[a]	2,151	3,372	1,484	2,031	1,703
Lower 95% CI of mean[a]	4,059	6,032	3,471	3,907	3,273
Upper 95% CI of mean[a]	7,136	10,856	5,594	6,812	8,693

Post-hoc pairwise comparison	P-value	Adjusted P	Summary[b]
HD1 - HD2	0.084	0.169	ns
HD1 - HD3	0.169	0.254	ns
HD1 - HD4	0.695	0.695	ns
HD2 - HD3	0.039	0.169	ns
HD2 - HD4	0.067	0.169	ns
HD3 - HD4	0.245	0.294	ns

[a]copies per ml plasma

[b]ns, not significant.

Supplementary Table S3.6 | Plasma *GAPDH* cDNA concentration summaries and statistics as measured by ddPCR.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	94	189	210	382.5	153.5
Median[a]	132.5	371.5	259	411.5	327.6
Maximum[a]	266	489	372	481	420.3
Mean[a]	153.5	376	279.2	420.3	307.3
Std. Deviation[a]	55.06	89	46.99	36.26	118.2
Lower 95% CI of mean[a]	114.1	312.3	245.6	394.3	119.1
Upper 95% CI of mean[a]	192.8	439.7	312.8	446.2	495.4

Post-hoc pairwise comparison	P-value	Adjusted P	Summary[b]
HD1 - HD2	0.004	0.012	*
HD1 - HD3	0.008	0.012	*
HD1 - HD4	0.003	0.012	*
HD2 - HD3	0.026	0.032	*
HD2 - HD4	0.121	0.121	ns
HD3 - HD4	0.007	0.012	*

[a]counts

[b] *, $P < 0.05$; ns, not significant.

Supplementary Table S3.7 | Plasma CD81+ EV count summaries and statistics as measured by flow cytometry.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	302.5	205	72	3921	177.1
Median[a]	572.8	397	112	5004	493.6
Maximum[a]	1043	570	742	5739	4843
Mean[a]	592.5	394.7	177.1	4843	1502
Std. Deviation[a]	198.2	114	201.1	577.5	2234
Lower 95% CI of mean[a]	450.7	313.1	33.26	4429	-2053
Upper 95% CI of mean[a]	734.2	476.3	320.9	5256	5056

Post-hoc pairwise comparison	P-value	Adjusted P	Summary[b]
HD1 - HD2	0.034	0.034	*
HD1 - HD3	0.010	0.014	*
HD1 - HD4	0.003	0.006	**
HD2 - HD3	0.033	0.034	*
HD2 - HD4	0.003	0.006	**
HD3 - HD4	0.003	0.006	**

[a]counts

[b] *, $P < 0.05$; **, $P < 0.01$; ns, not significant.

Supplementary Table S3.8 | Plasma CD63+ EV count summaries and statistics as measured by flow cytometry.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	2340	2822	678	3837	2620
Median[a]	3200	3177	1485	5082	3183
Maximum[a]	3619	3766	14028	6782	5075
Mean[a]	3140	3226	2620	5075	3515
Std. Deviation[a]	353.9	273.5	4038	772.3	1074
Lower 95% CI of mean[a]	2886	3030	-268.1	4523	1807
Upper 95% CI of mean[a]	3393	3421	5509	5628	5224

[a]counts

Supplementary Table S3.9 | Plasma CD41+ EV count summaries and statistics as measured by flow cytometry.

	HD1	HD2	HD3	HD4	Overall
Number of values	10	10	10	10	4
Minimum[a]	1022	363	512	439.5	487.3
Median[a]	1471	469.5	957	818	1143
Maximum[a]	2445	642	5340	1880	1485
Mean[a]	1485	487.3	1362	924.2	1065
Std. Deviation[a]	460.1	103.1	1429	437.8	453.9
Lower 95% CI of mean[a]	1156	413.5	339.5	611	342.3
Upper 95% CI of mean[a]	1814	561.1	2384	1237	1787

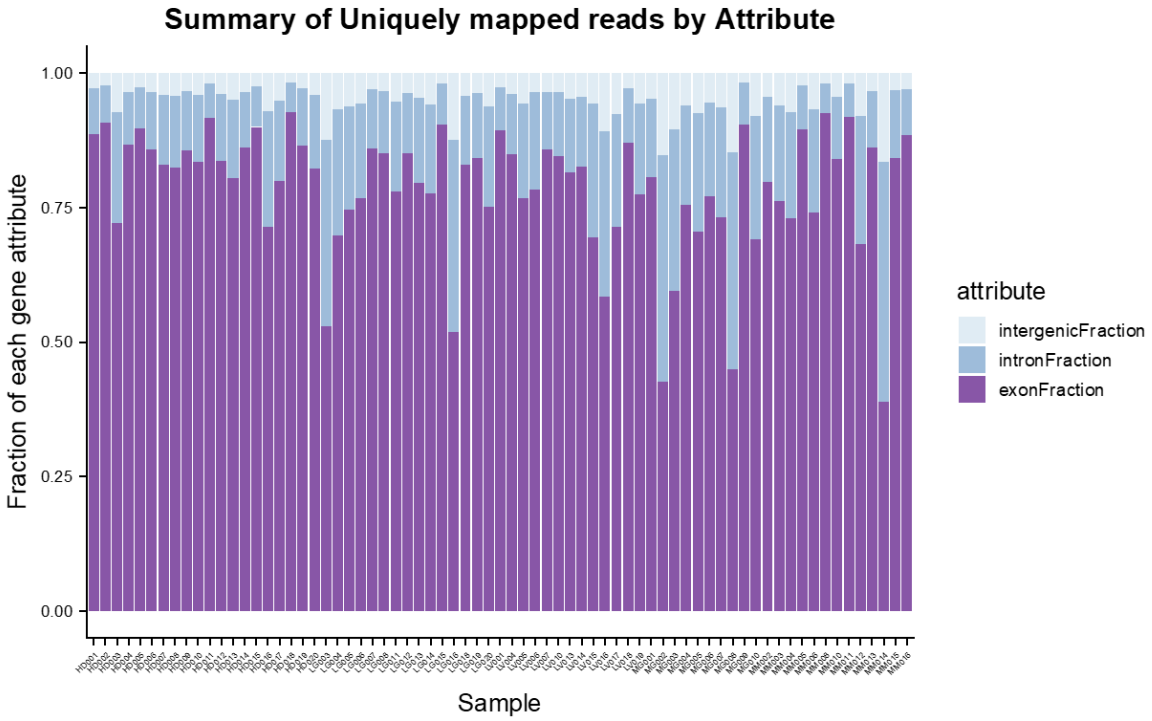
[a]counts

Supplementary Table S3.10 | Plasma CD9+ EV count summaries and statistics as measured by flow cytometry.

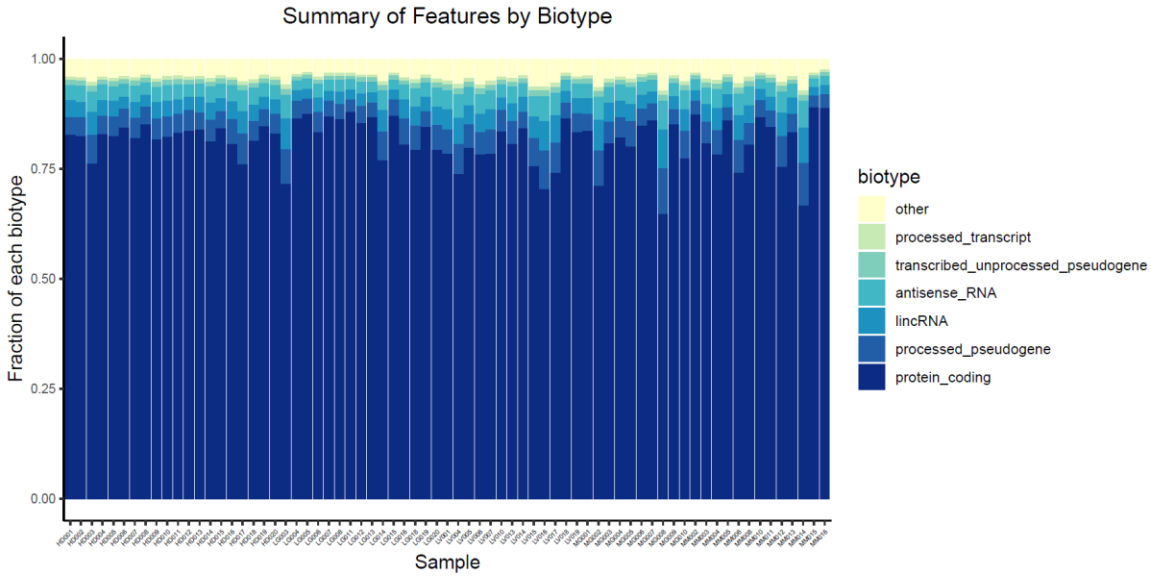
PP_ID	RNA Extraction	Library Preparation	Total Number of Reads	Number of Unique Reads	Exon Fraction	Intron Fraction	Intergenic Fraction	Protein Coding Fraction
PP002	batch 1	batch 1	33192487	9477957	0.93	0.05	0.02	0.81
PP003	batch 1	batch 1	29467052	4214981	0.86	0.11	0.03	0.82
PP004	batch 1	batch 1	29574675	3825380	0.82	0.13	0.04	0.85
PP005	batch 1	batch 1	33768635	4088109	0.86	0.1	0.03	0.83
PP007	batch 1	batch 1	31698177	1171491	0.84	0.13	0.03	0.89
PP008	batch 1	batch 1	30079806	10677465	0.74	0.19	0.07	0.74
PP010	batch 2	batch 1	38503355	3863232	0.76	0.18	0.06	0.81
PP011	batch 2	batch 1	31034252	4395327	0.86	0.11	0.04	0.84
PP012	batch 2	batch 1	35799309	4766786	0.92	0.07	0.02	0.83
PP014	batch 2	batch 1	27630599	3280679	0.85	0.11	0.04	0.86
PP015	batch 2	batch 1	30093097	6852599	0.84	0.12	0.04	0.87
PP016	batch 2	batch 1	29890017	6853766	0.86	0.1	0.04	0.81
PP017	batch 3	batch 1	34646834	6092611	0.69	0.23	0.08	0.77
PP018	batch 3	batch 1	35833774	3469637	0.59	0.3	0.1	0.81
PP019	batch 3	batch 1	36506970	4711445	0.72	0.21	0.07	0.76
PP020	batch 3	batch 1	34201040	5321020	0.8	0.15	0.05	0.76
PP026	batch 4	batch 1	40104631	3473624	0.83	0.13	0.05	0.84
PP027	batch 4	batch 1	34476121	17271358	0.85	0.11	0.04	0.74
PP028	batch 4	batch 1	33168567	10433701	0.78	0.17	0.06	0.77
PP029	batch 4	batch 1	30273014	3882069	0.9	0.08	0.02	0.84
PP031	batch 4	batch 1	35139077	9791729	0.7	0.25	0.06	0.76
PP032	batch 4	batch 1	31904022	4016176	0.87	0.1	0.04	0.83
PP034	batch 4	batch 2	36949075	2421501	0.8	0.16	0.05	0.87
PP035	batch 4	batch 2	30996985	13700958	0.86	0.11	0.04	0.78
PP036	batch 4	batch 2	36919622	7006298	0.71	0.22	0.07	0.8
PP038	batch 4	batch 2	33199864	2412155	0.84	0.12	0.04	0.85
PP039	batch 4	batch 2	35675985	5024440	0.77	0.18	0.06	0.83
PP041	batch 5	batch 2	35773928	6240377	0.83	0.13	0.04	0.82
PP042	batch 5	batch 2	35650935	3999501	0.86	0.11	0.03	0.87
PP043	batch 5	batch 2	31753026	7231832	0.71	0.22	0.07	0.81
PP046	batch 5	batch 2	30298636	4380430	0.84	0.13	0.04	0.84
PP047	batch 5	batch 2	31468616	13876029	0.93	0.05	0.02	0.81
PP048	batch 5	batch 2	32642636	8534075	0.82	0.14	0.05	0.81
PP049	batch 6	batch 2	37084139	9845748	0.88	0.08	0.03	0.89
PP050	batch 6	batch 2	40532989	5076857	0.87	0.1	0.03	0.87
PP052	batch 6	batch 2	34404394	6241844	0.9	0.08	0.02	0.87
PP055	batch 6	batch 2	34851176	15323436	0.91	0.07	0.02	0.83
PP056	batch 6	batch 2	41922873	4579812	0.85	0.11	0.03	0.86
PP058	batch 7	batch 2	36866771	5758608	0.9	0.08	0.02	0.85
PP060	batch 7	batch 2	37126236	4991733	0.75	0.19	0.06	0.82
PP061	batch 7	batch 2	37509562	13009859	0.78	0.18	0.03	0.78
PP062	batch 7	batch 2	36883791	21874141	0.9	0.08	0.03	0.83
PP063	batch 7	batch 2	43184305	5024575	0.75	0.19	0.06	0.88
PP067	batch 7	batch 3	38201238	4939376	0.7	0.23	0.07	0.87
PP073	batch 8	batch 3	43566479	2482378	0.8	0.16	0.04	0.87
PP074	batch 8	batch 3	34208260	5631056	0.92	0.06	0.02	0.85
PP075	batch 8	batch 3	31957614	5211761	0.84	0.12	0.04	0.84
PP077	batch 8	batch 3	38776982	4057998	0.52	0.36	0.12	0.81
PP078	batch 8	batch 3	36596920	14457187	0.89	0.08	0.03	0.78
PP084	batch 9	batch 3	52307262	10415932	0.43	0.42	0.15	0.71
PP087	batch 9	batch 3	34570617	2699261	0.8	0.15	0.05	0.84
PP091	batch 10	batch 3	38121200	3181309	0.82	0.14	0.04	0.83
PP097	batch 11	batch 4	27998299	3867357	0.81	0.15	0.05	0.84
PP098	batch 11	batch 4	27748334	14243418	0.68	0.24	0.08	0.75
PP099	batch 11	batch 4	28514250	6664874	0.77	0.18	0.06	0.8
PP101	batch 11	batch 4	27393529	15434647	0.53	0.35	0.12	0.72
PP102	batch 11	batch 4	28850027	7551982	0.89	0.09	0.03	0.83
PP103	batch 11	batch 4	27808286	14620805	0.45	0.4	0.15	0.65
PP105	batch 11	batch 4	30350083	8031052	0.83	0.13	0.04	0.82
PP107	batch 12	batch 4	28853047	9469998	0.75	0.19	0.06	0.79
PP109	batch 12	batch 4	28245277	14323721	0.71	0.21	0.08	0.74
PP111	batch 12	batch 4	33349408	7413209	0.77	0.17	0.06	0.83
PP112	batch 12	batch 4	30195039	8694949	0.83	0.13	0.04	0.79
PP114	batch 12	batch 4	33053778	2478571	0.73	0.2	0.06	0.86
PP115	batch 12	batch 4	32223434	1498604	0.78	0.17	0.05	0.88
PP116	batch 12	batch 4	31973526	4755722	0.87	0.11	0.03	0.85

Supplementary Table S4.1 | Summary of input reads, unique reads, exon fraction, intron fraction, intergenic fraction, and protein coding fraction.

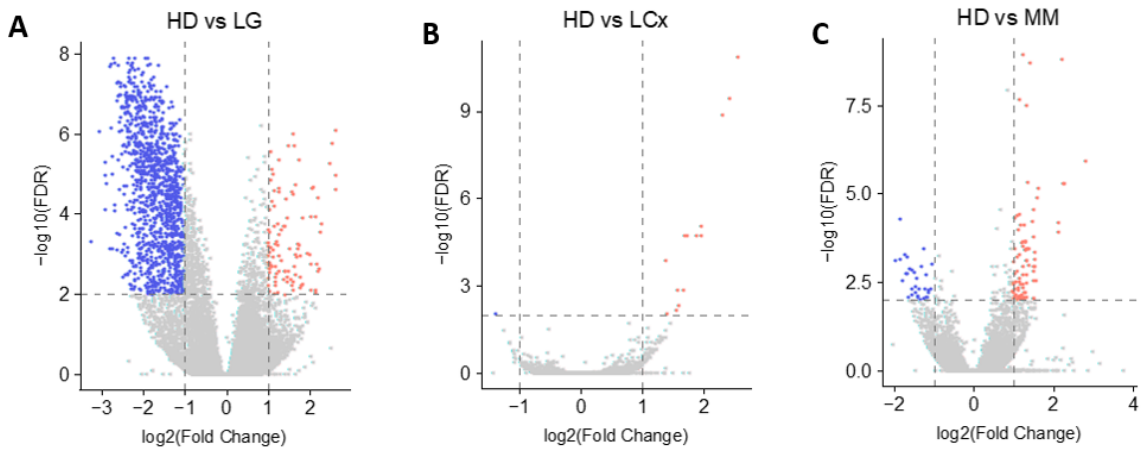
Supplementary Figure S4.1 | Distribution of sequencing reads across all 71 samples.



Supplementary Figure S4.2 | Distribution of exon/intro and intergenic fractions across all 71 samples.

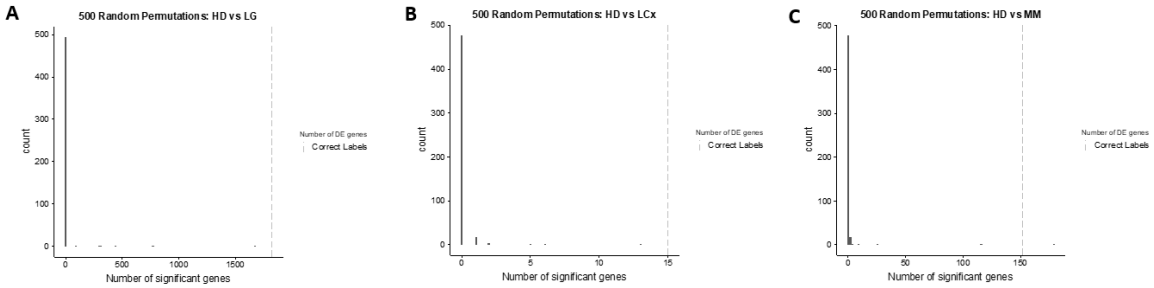


Supplementary Figure S4.3 | Coverage of the transcriptome across all 71 samples



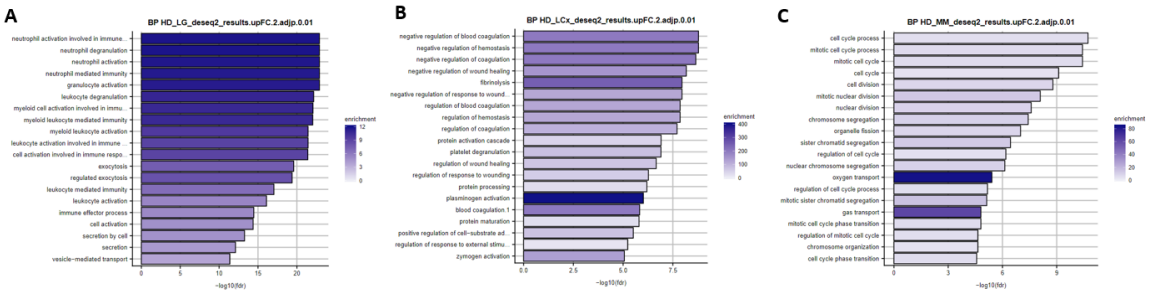
Supplementary Figure S4.4 | Volcano plots from cfRNA pairwise cohorts

Volcano plots between false discovery rate and fold changes for all genes of pairwise comparison between healthy donors (HD) and lung cancer (LG, panel A), liver cancer (LCx, panel B) and multiple myeloma (MM, panel C) analyzed by DESeq2.



Supplementary Figure S4.5 | Differential gene expression permutation tests

Histograms of number of significant genes differentiating two groups from random permutation between samples across healthy donors and lung cancer (A), liver cancer (B) or multiple myeloma (C). Differential expression analysis was performed using DESeq2 with Wald test and padj value cut off at 0.01.

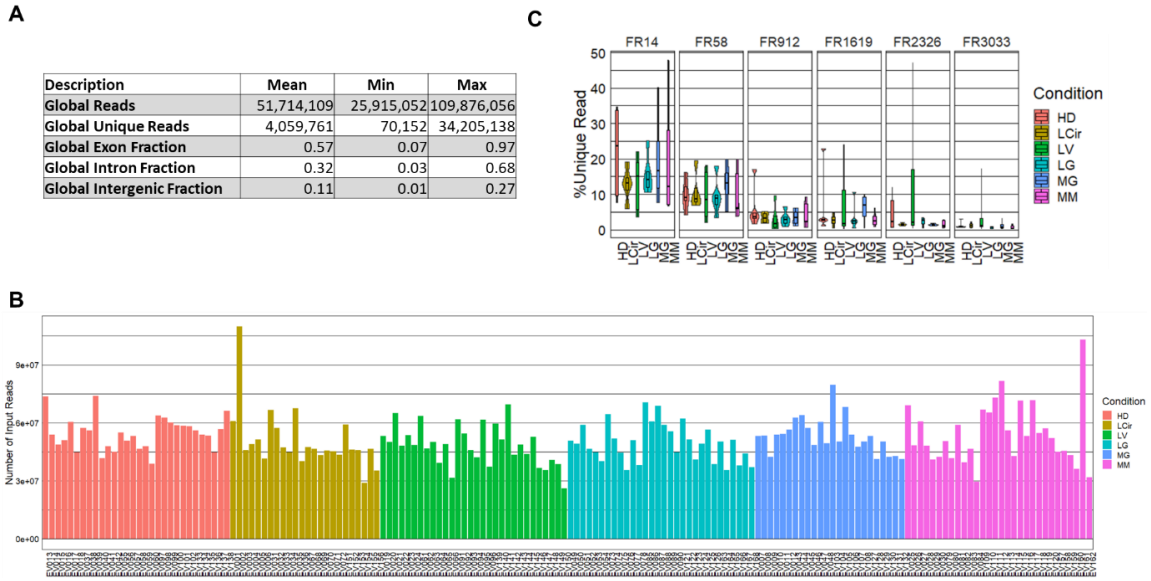


Supplementary Figure S4.6 | cfRNA Gene Ontology

Gene Ontology analysis show the enrichment of biological processes for significant gene panels identified by DESeq2 analysis for pairwise comparison between healthy and lung cancer (A), liver cancer (B) and multiple myeloma (C).

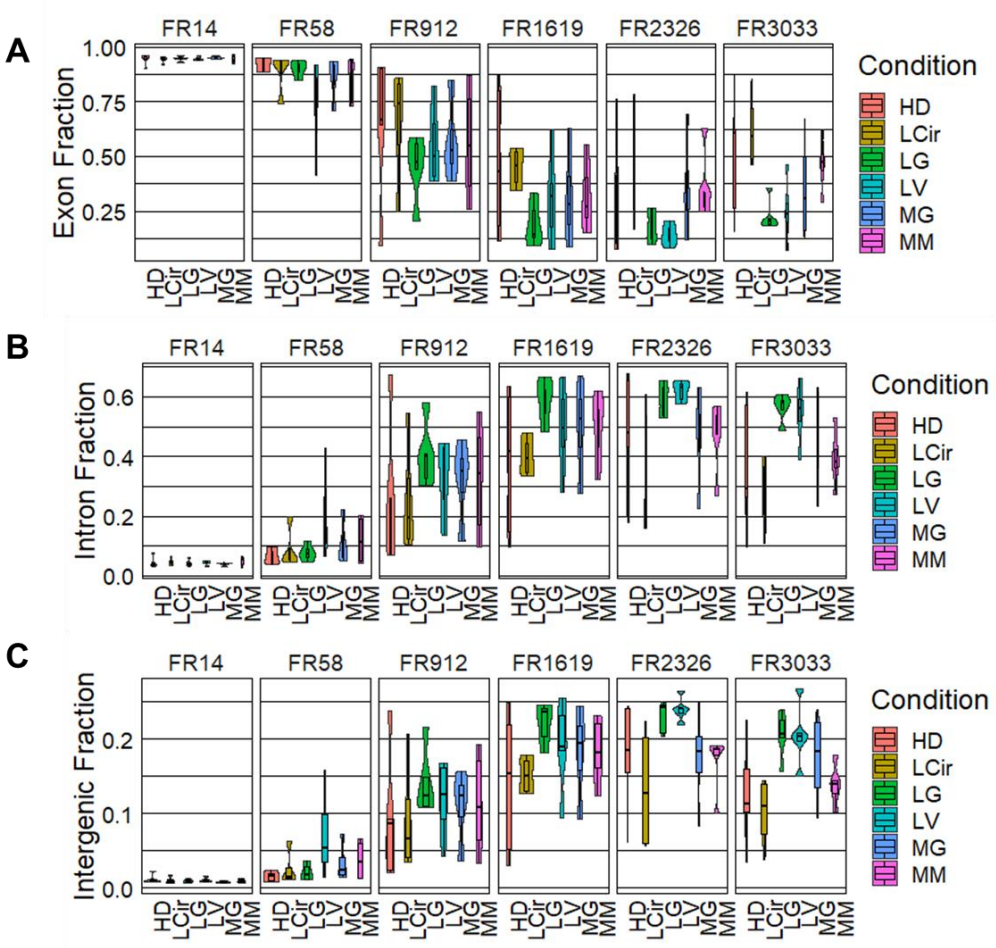
A

Description	Mean	Min	Max
Global Reads	51,714,109	25,915,052	109,876,056
Global Unique Reads	4,059,761	70,152	34,205,138
Global Exon Fraction	0.57	0.07	0.97
Global Intron Fraction	0.32	0.03	0.68
Global Intergenic Fraction	0.11	0.01	0.27



Supplementary Figure S5.1 | Description of input reads, unique reads, exon, intron and intergenic fraction

(A) Table of description of global reads, unique reads, exon fraction, intron fraction, and intergenic fraction across 168 sequencing sample. The average, minimum and maximum values are shown from RNA-seq quality control package (RSeQc). (B) Bar graph of number of input reads across 168 sequencing sample colored by conditions. (C) The violin plot of unique reads percentage grouped by each condition across fractions.



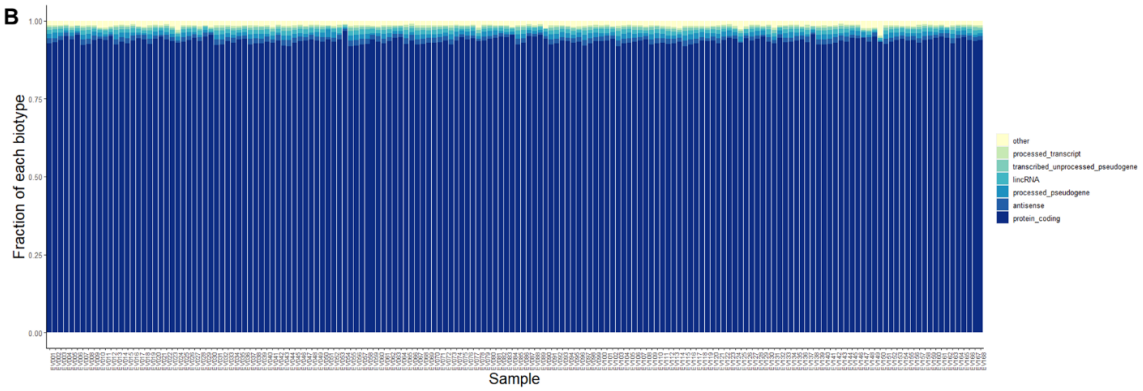
Supplementary Figure S5.2 | Distribution of exon, intron, and intergenic fractions across fractions

Violin plots across plasma fractions (FR14, FR58, FR912, FR1619, FR2326, and FR3033) grouped by each conditions showing the respective fraction of (A) Exons, (B) Introns, (C) Intergenic reads.

A

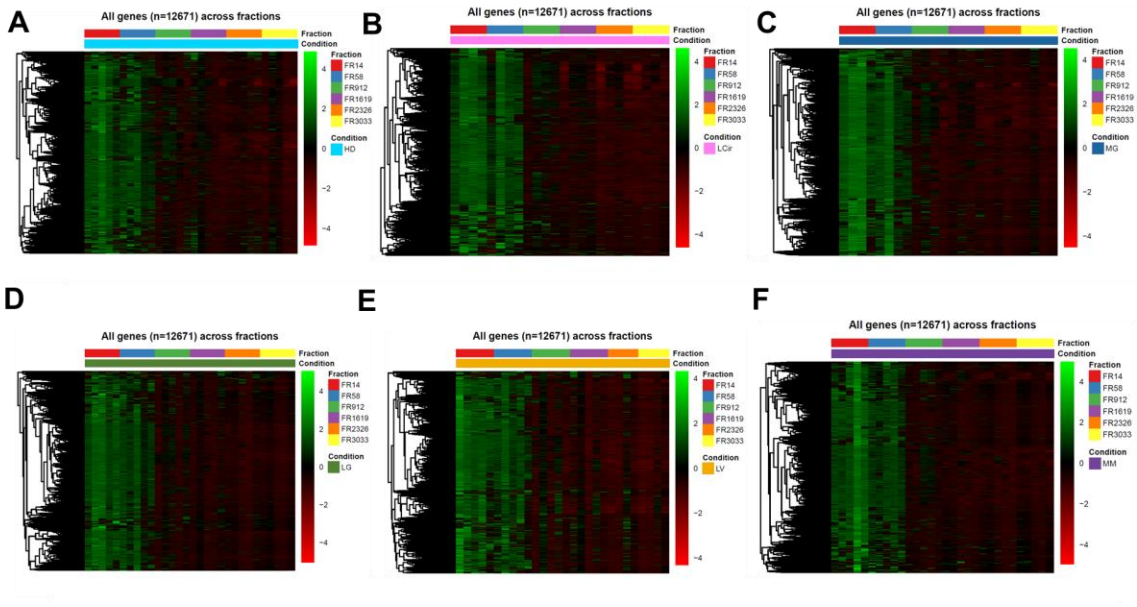
Biotype	Mean	Min	Max
protein_coding	93.50%	91.75%	96.66%
transcribed_unprocessed_pseudogene	0.71%	0.19%	1.21%
processed_transcript	0.52%	0.13%	0.92%
processed_pseudogene	1.23%	0.38%	2.40%
lincRNA	1.14%	0.40%	1.67%
antisense	1.29%	0.16%	1.90%
other	1.62%	0.89%	4.76%

B



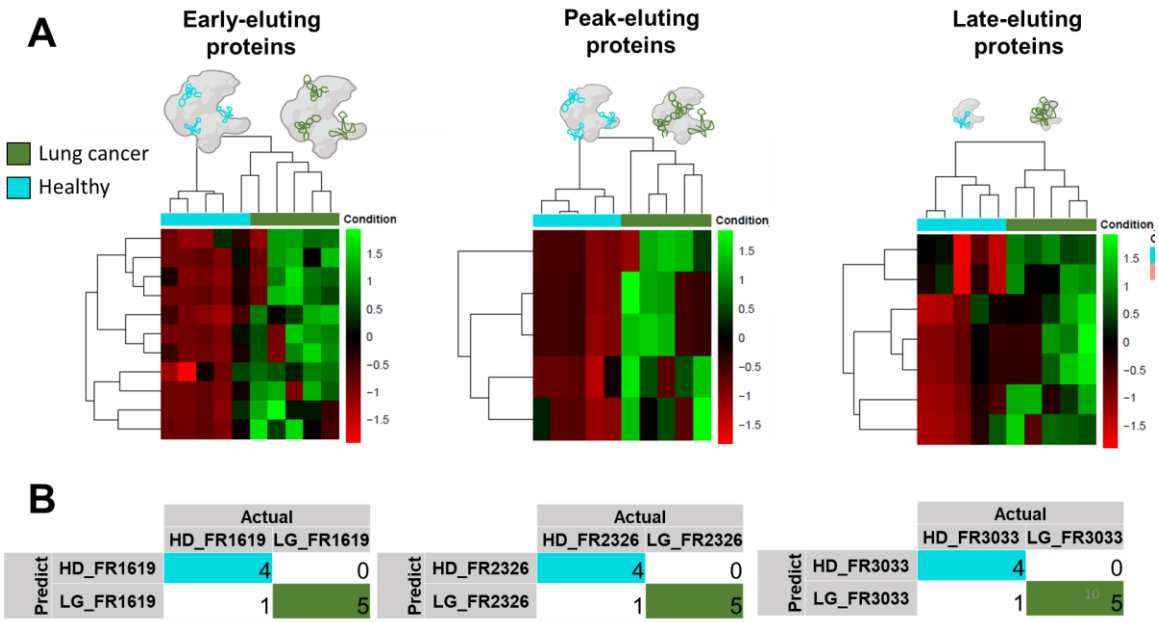
Supplementary Figure S5.3 | Summary of proportion of transcript types across all samples

(A) Table of biotype categories including protein coding, transcribed unprocessed pseudogene, processed transcript, processed pseudogene, lincRNA, antisense, and others. The average, minimum and maximum values are shown. (B) A stack column representing fraction of each biotype across 168 samples.



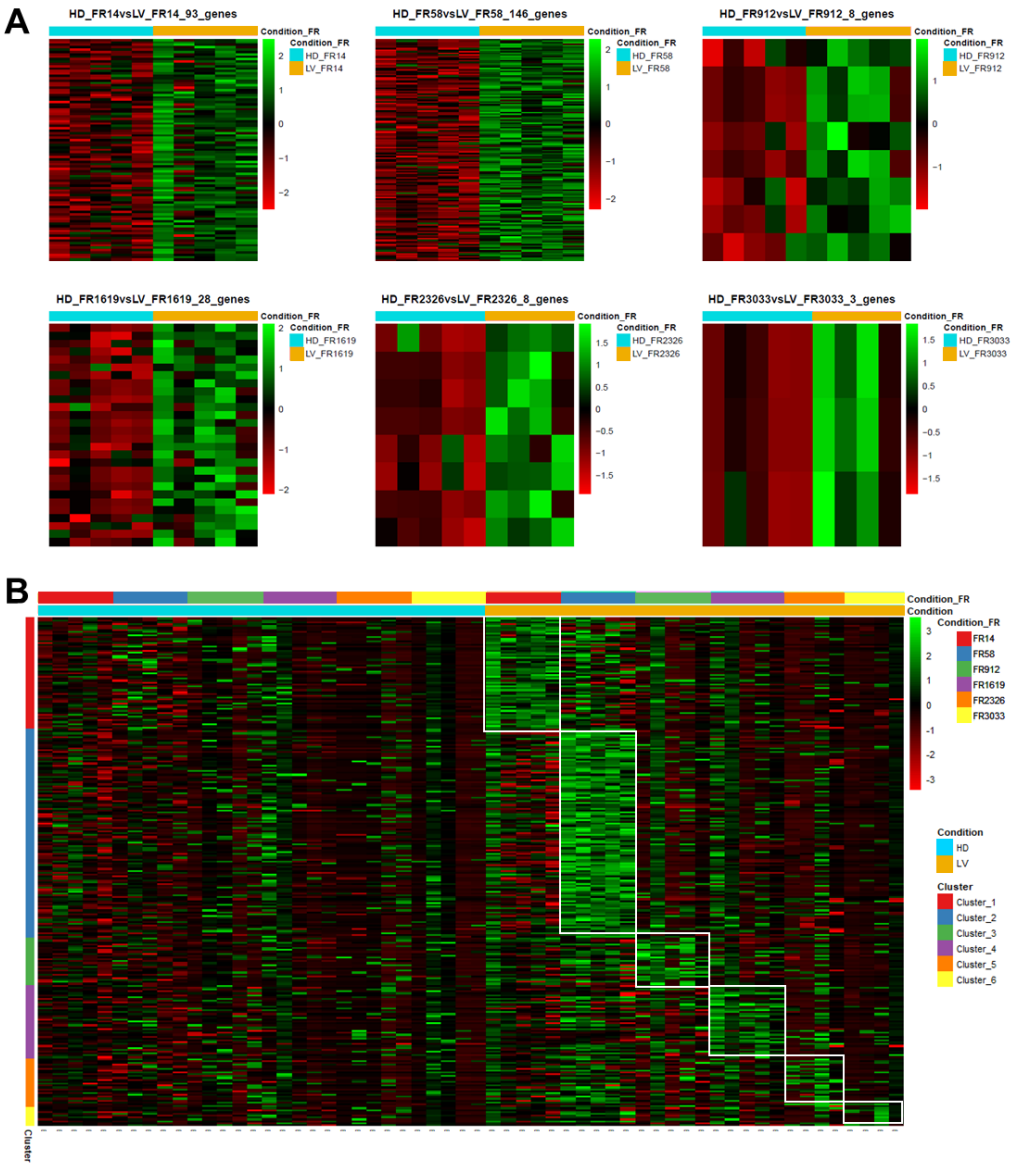
Supplementary Figure S5.4 | Transcriptomic analysis of EVs and non-vesicles per condition

Heatmap expression of all genes across all fractions from (A) healthy sample, (B) liver cirrhosis, (C) monoclonal gammopathy of undetermined significance, (D) lung cancer, (E) liver cancer, and (F) multiple myeloma. HD: healthy, LCir: liver cirrhosis, MG: monoclonal gammopathy of undetermined significance, LG: lung cancer, LV: liver cancer, and MM: multiple myeloma.



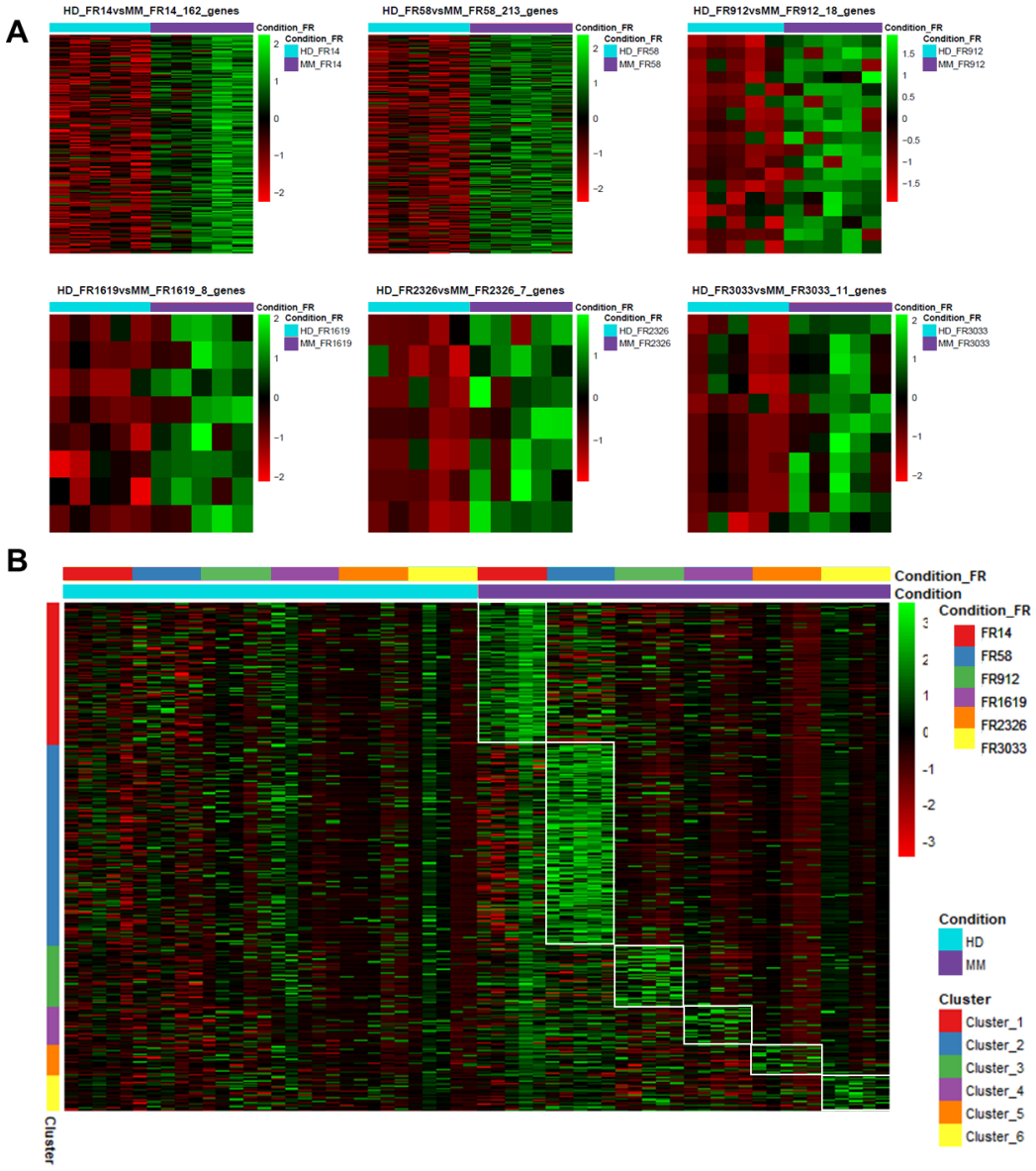
Supplementary Figure S5.5 | Lung cancer associated genes packaged in protein enriched fraction

(A) Heatmap of log counts of lung cancer differentially expressed genes within individual fractions enriched in early-, peak-, and late-eluting proteins were compared between lung cancer and healthy. Differentially expressed genes which showed statistical significance (student's t test, p-value < 0.05) were used. (B) Leave-one-out cross validation to test linear discriminant analysis algorithm accuracy for classification using DE gene sets identified from early-, peak-, and late-eluting fraction up-regulated in lung cancer compared to healthy.



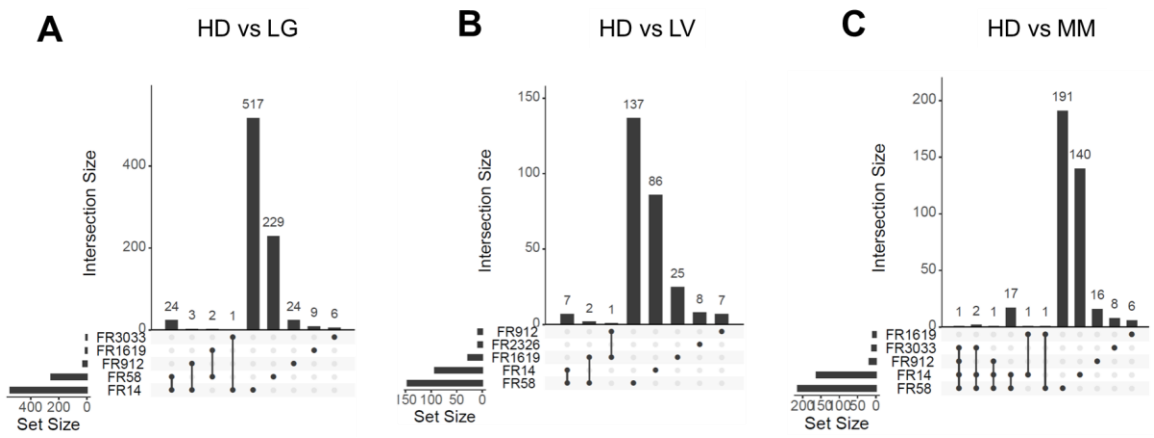
Supplementary Figure S5.6 | Liver cancer associated genes across fractions in human plasma

(A) Heatmap of log counts of liver cancer differentially expressed genes within individual fractions between each cancer type and healthy. Differentially expressed genes which showed statistical significance (student's t test, p-value < 0.05) were used. (B) Heatmap of gene expression in liver cancer relative to healthy across fractions. A total of 270 significantly differentially expressed genes were used. Clusters were assigned to genes corresponding to their enriched fraction based on log 2 fold changes (FR14, FR58, FR912, FR1619, FR2326, FR3033 as clusters 1-6 respectively).



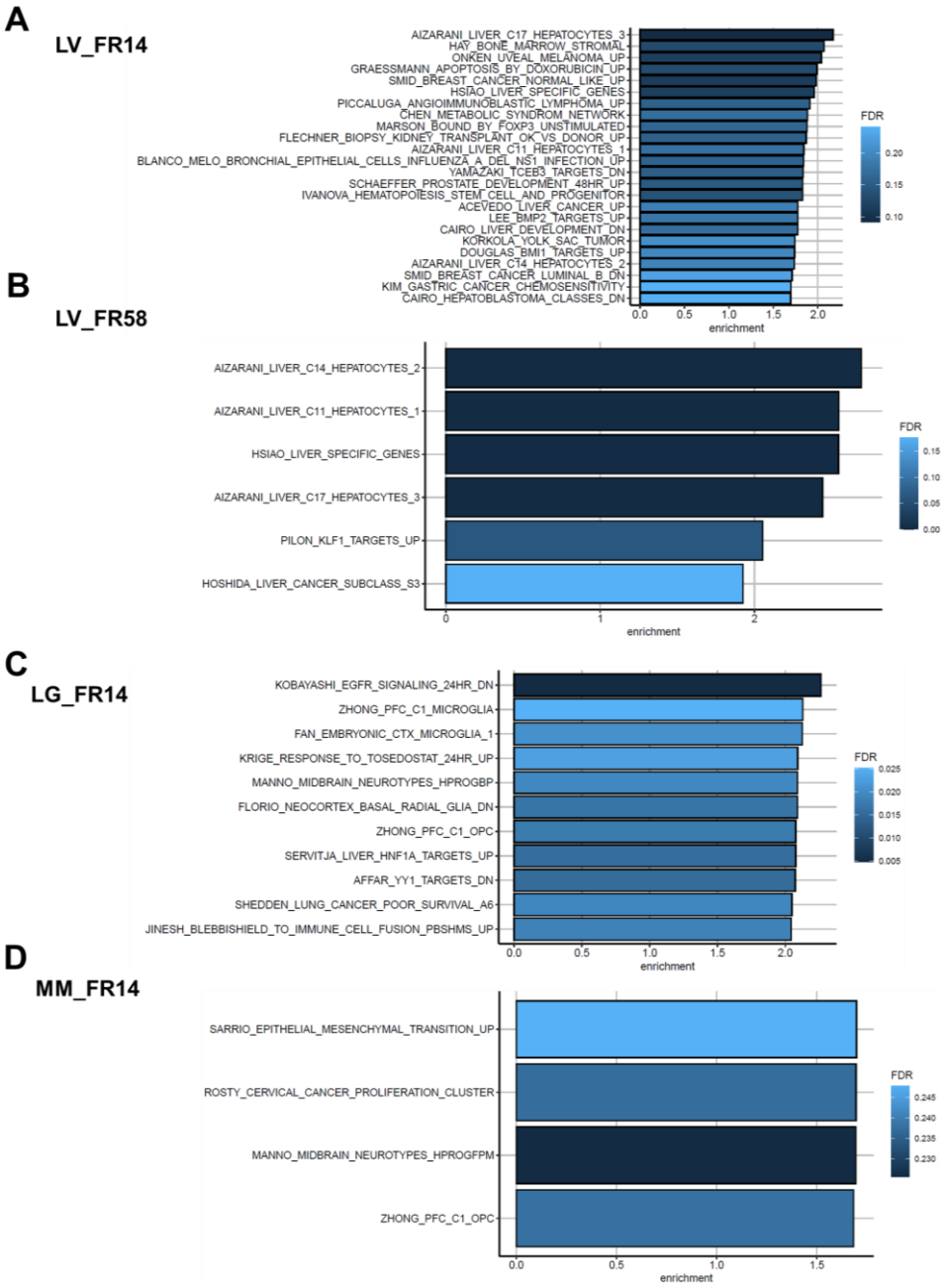
Supplementary Figure S5.7 | Multiple myeloma associated genes across fractions in human plasma

(A) Heatmap of log counts of lung cancer differentially expressed genes within individual fractions between each multiple myeloma and healthy. Differentially expressed genes which showed statistical significance (student's t test, p-value < 0.05) were used. (B) Heatmap of gene expression in multiple myeloma relative to healthy across fractions. A total of 381 significantly differentially expressed genes were used. Clusters were assigned to genes corresponding to their enriched fraction based on log 2 fold changes (FR14, FR58, FR912, FR1619, FR2326, FR3033 as clusters 1-6 respectively).



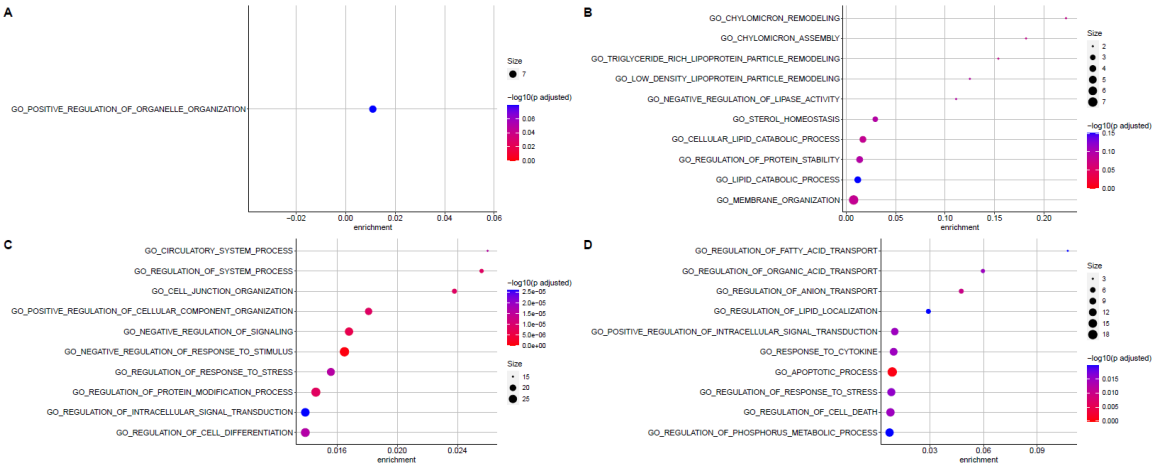
Supplementary Figure S5.8 | Intersection of cancer distinguishing genes across plasma fractions

Upset plot for cancer distinguishing genes identified in individual fraction for (A) lung cancer, (B) liver cancer, and (C) multiple myeloma



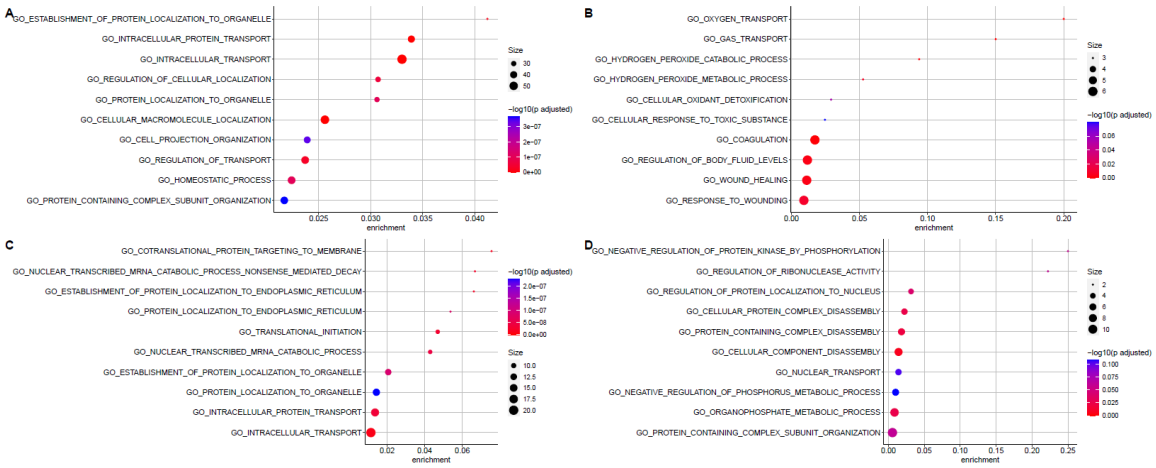
Supplementary Figure S5.9 | Gene set enrichment for cancer distinguishing genes

Gene set enrichment analysis (GSEA) was performed on cancer distinguishing genes enriched in specific fraction. GSEA was performed from Molecular signatures database (MSigDB, https://www.broadinstitute.org/gsea/msigdb/) using C2:CGP chemical and genetic perturbation for (A) liver cancer genes enriched in fraction 1-4 (FR14), (B) liver cancer genes enriched in fraction 5-8 (FR58), (C) lung cancer genes enriched in FR14, and (D) multiple myeloma genes enriched in FR14.



Supplementary Figure S5.10 | Gene set enrichment analysis associated with healthy, liver cirrhosis, and liver cancer comparisons

Gene set enrichment analysis (GSEA) was performed on specific cfRNA signatures associated with liver cirrhosis and liver cancer. GSEA was performed from Molecular signatures database (MSigDB, https://www.broadinstitute.org/gsea/msigdb/) using C5:BP derived from biological process ontology for (A) healthy upregulated genes, (B) liver cancer upregulated genes, (C) healthy and liver cirrhosis upregulated genes, and (D) liver cirrhosis and liver cancer upregulated genes.



Supplementary Figure S5.11 | Gene ontology analysis associated with healthy, MGUS, and multiple myeloma comparisons

Gene set enrichment analysis (GSEA) was performed on specific cfRNA signatures associated with monoclonal gammopathy of undetermined significance (MGUS) and multiple myeloma. GSEA was performed from Molecular signatures database (MSigDB, https://www.broadinstitute.org/gsea/msigdb/) using C5:BP derived from biological process ontology for (A) healthy upregulated genes, (B) multiple myeloma upregulated genes, (C) healthy and MGUS upregulated genes, and (D) MGUS and multiple myeloma upregulated genes.

References

1. Pan, B.-T. and R.M. Johnstone, *Fate of the transferrin receptor during maturation of sheep reticulocytes in vitro: Selective externalization of the receptor.* Cell, 1983. **33**(3): p. 967-978.

2. Kalluri, R. and V.S. LeBleu, *The biology, function, and biomedical applications of exosomes.* Science, 2020. **367**(6478): p. eaau6977.

3. van Niel, G., et al., *Exosomes: A Common Pathway for a Specialized Function.* The Journal of Biochemistry, 2006. **140**(1): p. 13-21.

4. Zhang, H. and D. Lyden, *Asymmetric-flow field-flow fractionation technology for exomere and small extracellular vesicle separation and characterization.* Nature protocols, 2019. **14**(4): p. 1027-1053.

5. Pegtel, D.M. and S.J. Gould, *Exosomes.* Annual Review of Biochemistry, 2019. **88**(1): p. 487-514.

6. Hemler, M.E., *Tetraspanin proteins mediate cellular penetration, invasion, and fusion events and define a novel type of membrane microdomain.* Annual review of cell and developmental biology, 2003. **19**(1): p. 397-422.

7. Kalluri, R., *The biology and function of exosomes in cancer.* The Journal of clinical investigation, 2016. **126**(4): p. 1208-1215.

8. Steinbichler, T.B., et al., *The role of exosomes in cancer metastasis.* Semin Cancer Biol, 2017. **44**: p. 170-181.

9. Camussi, G., et al., *Exosomes/microvesicles as a mechanism of cell-to-cell communication.* Kidney International, 2010. **78**(9): p. 838-848.

10. Hoshino, A., et al., *Tumour exosome integrins determine organotropic metastasis.* Nature, 2015. **527**(7578): p. 329-335.

11. Demory Beckler, M., et al., *Proteomic Analysis of Exosomes from Mutant KRAS Colon Cancer Cells Identifies Intercellular Transfer of Mutant KRAS*.* Molecular & Cellular Proteomics, 2013. **12**(2): p. 343-355.

12. Choi, D., et al., *The impact of oncogenic EGFRvIII on the proteome of extracellular vesicles released from glioblastoma cells.* Molecular & Cellular Proteomics, 2018. **17**(10): p. 1948-1964.

13. Gehrmann, U., et al., *Harnessing the exosome-induced immune response for cancer immunotherapy.* Seminars in Cancer Biology, 2014. **28**: p. 58-67.

14. Kurywchak, P., J. Tavormina, and R. Kalluri, *The emerging roles of exosomes in the modulation of immune responses in cancer.* Genome Medicine, 2018. **10**(1): p. 23.

15. Lacroix, R., et al., *Impact of pre-analytical parameters on the measurement of circulating microparticles: towards standardization of protocol.* J Thromb Haemost, 2012. **10**(3): p. 437-46.

16. Lacroix, R., et al., *Standardization of pre-analytical variables in plasma microparticle determination: results of the International Society on Thrombosis and Haemostasis SSC Collaborative workshop.* Journal of thrombosis and haemostasis : JTH, 2013: p. 10.1111/jth.12207.

17. Yuana, Y., R.M. Bertina, and S. Osanto, *Pre-analytical and analytical issues in the analysis of blood microparticles.* Thromb Haemost, 2011. **105**(3): p. 396-408.

18. Yuana, Y., et al., *Handling and storage of human body fluids for analysis of extracellular vesicles.* Journal of extracellular vesicles, 2015. **4**: p. 29260-29260.

19. Gurunathan, S., et al., *Review of the Isolation, Characterization, Biological Function, and Multifarious Therapeutic Approaches of Exosomes.* Cells, 2019. **8**(4): p. 307.

20. Brennan, K., et al., *A comparison of methods for the isolation and separation of extracellular vesicles from protein and lipid particles in human serum.* Scientific Reports, 2020. **10**(1): p. 1039.

21. Konoshenko, M.Y., et al., *Isolation of Extracellular Vesicles: General Methodologies and Latest Trends.* BioMed Research International, 2018. **2018**: p. 8545347.

22. Baranyai, T., et al., *Isolation of Exosomes from Blood Plasma: Qualitative and Quantitative Comparison of Ultracentrifugation and Size Exclusion Chromatography Methods.* PLOS ONE, 2015. **10**(12): p. e0145686.

23. Benedikter, B.J., et al., *Ultrafiltration combined with size exclusion chromatography efficiently isolates extracellular vesicles from cell culture media for compositional and functional studies.* Scientific Reports, 2017. **7**(1): p. 15297.

24. Ramirez, M.I., et al., *Technical challenges of working with extracellular vesicles.* Nanoscale, 2018. **10**(3): p. 881-906.

25. Willis, G.R., S. Kourembanas, and S.A. Mitsialis, *Toward Exosome-Based Therapeutics: Isolation, Heterogeneity, and Fit-for-Purpose Potency.* Frontiers in Cardiovascular Medicine, 2017. **4**(63).

26. Corso, G., et al., *Reproducible and scalable purification of extracellular vesicles using combined bind-elute and size exclusion chromatography.* Scientific Reports, 2017. **7**(1): p. 11561.

27. Böing, A.N., et al., *Single-step isolation of extracellular vesicles by size-exclusion chromatography.* Journal of extracellular vesicles, 2014. **3**(1): p. 23430.

28. Zarovni, N., et al., *Integrated isolation and quantitative analysis of exosome shuttled proteins and nucleic acids using immunocapture approaches.* Methods, 2015. **87**: p. 46-58.

29. Jørgensen, M., et al., *Extracellular Vesicle (EV) Array: microarray capturing of exosomes and other extracellular vesicles for multiplexed phenotyping.* Journal of extracellular vesicles, 2013. **2**(1): p. 20920.

30. Zhang, H., et al., *Identification of distinct nanoparticles and subsets of extracellular vesicles by asymmetric flow field-flow fractionation.* Nature cell biology, 2018. **20**(3): p. 332-343.

31. Zhang, H. and D. Lyden, *Asymmetric-flow field-flow fractionation technology for exomere and small extracellular vesicle separation and characterization.* Nat Protoc, 2019. **14**(4): p. 1027-1053.

32. Val, S., et al., *Purification and characterization of microRNAs within middle ear fluid exosomes: implication in otitis media pathophysiology.* Pediatric Research, 2017. **81**(6): p. 911-918.

33. Stetefeld, J., S.A. McKenna, and T.R. Patel, *Dynamic light scattering: a practical guide and applications in biomedical sciences.* Biophysical reviews, 2016. **8**(4): p. 409-427.

34. Jung, M.K. and J.Y. Mun, *Sample Preparation and Imaging of Exosomes by Transmission Electron Microscopy.* Journal of visualized experiments : JoVE, 2018(131): p. 56482.

35. Chuo, S.T.-Y., J.C.-Y. Chien, and C.P.-K. Lai, *Imaging extracellular vesicles: current and emerging methods.* Journal of Biomedical Science, 2018. **25**(1): p. 91.

36. Emelyanov, A., et al., *Cryo-electron microscopy of extracellular vesicles from cerebrospinal fluid.* PLoS One, 2020. **15**(1): p. e0227949.

37. Maas, S.L.N., J. De Vrij, and M.L.D. Broekman, *Quantification and size-profiling of extracellular vesicles using tunable resistive pulse sensing.* Journal of visualized experiments : JoVE, 2014(92): p. e51623-e51623.

38. Maas, S.L.N., M.L.D. Broekman, and J. de Vrij, *Tunable Resistive Pulse Sensing for the Characterization of Extracellular Vesicles,* in *Exosomes and Microvesicles:*

Methods and Protocols, A.F. Hill, Editor. 2017, Springer New York: New York, NY. p. 21-33.

39. Arroyo, J.D., et al., *Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma.* Proceedings of the National Academy of Sciences, 2011. **108**(12): p. 5003-5008.

40. Welsh, J.A., et al., *FCMPASS Software Aids Extracellular Vesicle Light Scatter Standardization.* Cytometry Part A, 2020. **97**(6): p. 569-581.

41. Van Der POL, E., et al., *Single vs. swarm detection of microparticles and exosomes by flow cytometry.* Journal of Thrombosis and Haemostasis, 2012. **10**(5): p. 919-930.

42. Vembadi, A., A. Menachery, and M.A. Qasaimeh, *Cell Cytometry: Review and Perspective on Biotechnological Advances.* Frontiers in Bioengineering and Biotechnology, 2019. **7**(147).

43. *Training & E-learning*. 2019 2019 [cited 2021 March 20]; Available from: https://www.bdbiosciences.com/us/support/s/itf_launch.

44. Gardiner, C., et al., *Measurement of refractive index by nanoparticle tracking analysis reveals heterogeneity in extracellular vesicles.* J Extracell Vesicles, 2014. **3**: p. 25361.

45. van der Pol, E., et al., *Refractive index determination of nanoparticles in suspension using nanoparticle tracking analysis.* Nano Lett, 2014. **14**(11): p. 6195-201.

46. Dragovic, R.A., et al., *Sizing and phenotyping of cellular vesicles using Nanoparticle Tracking Analysis.* Nanomedicine, 2011. **7**(6): p. 780-8.

47. Braeckmans, K., et al., *Sizing Nanomatter in Biological Fluids by Fluorescence Single Particle Tracking.* Nano Letters, 2010. **10**(11): p. 4435-4442.

48. Fattaccioli, J., et al., *Size and fluorescence measurements of individual droplets by flow cytometry.* Soft Matter, 2009. **5**(11): p. 2232-2238.

49. Welsh, J.A. and J.C. Jones, *Small Particle Fluorescence and Light Scatter Calibration Using FCMPASS Software.* Current Protocols in Cytometry, 2020. **94**(1): p. e79.

50. Welsh, J.A., J.C. Jones, and V.A. Tang, *Fluorescence and Light Scatter Calibration Allow Comparisons of Small Particle Data in Standard Units across Different Flow Cytometry Platforms and Detector Settings.* Cytometry Part A, 2020. **97**(6): p. 592-601.

51. Wang, L., et al., *Quantitating fluorescence intensity from fluorophores: practical use of MESF values.* Journal of research of the National Institute of Standards and Technology, 2002. **107**(4): p. 339.

52. Cox, J. and M. Mann, *Quantitative, high-resolution proteomics for data-driven systems biology.* Annu Rev Biochem, 2011. **80**: p. 273-99.

53. Zhu, W., J.W. Smith, and C.-M. Huang, *Mass Spectrometry-Based Label-Free Quantitative Proteomics.* Journal of Biomedicine and Biotechnology, 2010. **2010**: p. 840518.

54. Kito, K. and T. Ito, *Mass spectrometry-based approaches toward absolute quantitative proteomics.* Current genomics, 2008. **9**(4): p. 263-274.

55. Kowal, J., et al., *Proteomic comparison defines novel markers to characterize heterogeneous populations of extracellular vesicle subtypes.* Proceedings of the National Academy of Sciences, 2016. **113**(8): p. E968-E977.

56. Hoshino, A., et al., *Extracellular Vesicle and Particle Biomarkers Define Multiple Human Cancers.* Cell, 2020. **182**(4): p. 1044-1061.e18.

57. Everaert, C., et al., *Performance assessment of total RNA sequencing of human biofluids and extracellular vesicles.* Scientific Reports, 2019. **9**(1): p. 17574.

58. Das, S., et al., *The Extracellular RNA Communication Consortium: Establishing Foundational Knowledge and Technologies for Extracellular RNA Research.* Cell, 2019. **177**(2): p. 231-242.

59. Skog, J., et al., *Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers.* Nature Cell Biology, 2008. **10**(12): p. 1470-1476.

60. Murillo, O.D., et al., *exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids.* Cell, 2019. **177**(2): p. 463-477.e15.

61. Michell, D.L., et al., *Isolation of high-density lipoproteins for non-coding small RNA quantification.* JoVE (Journal of Visualized Experiments), 2016(117): p. e54488.

62. Arroyo, J.D., et al., *Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(12): p. 5003-5008.

63. McKenzie, A.J., et al., *KRAS-MEK Signaling Controls Ago2 Sorting into Exosomes.* Cell Rep, 2016. **15**(5): p. 978-987.

64. Mandel, P., *Les acides nucleiques du plasma sanguin chez 1 homme.* CR Seances Soc Biol Fil, 1948. **142**: p. 241-243.

65. Lo, K.-W., et al., *Analysis of cell-free Epstein-Barr virus-associated RNA in the plasma of patients with nasopharyngeal carcinoma.* Clinical chemistry, 1999. **45**(8): p. 1292-1294.

66. Kopreski, M.S., et al., *Detection of tumor messenger RNA in the serum of patients with malignant melanoma.* Clinical cancer research, 1999. **5**(8): p. 1961-1965.

67. Zuo, Z., et al., *BBCancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers.* Nucleic Acids Research, 2020. **48**(D1): p. D789-D796.

68. Mitchell, P.S., et al., *Circulating microRNAs as stable blood-based markers for cancer detection.* Proceedings of the National Academy of Sciences, 2008. **105**(30): p. 10513-10518.

69. de Planell-Saguer, M. and M.C. Rodicio, *Analytical aspects of microRNA in diagnostics: A review.* Analytica Chimica Acta, 2011. **699**(2): p. 134-152.

70. Sayeed, A., et al., *Profiling the circulating mRNA transcriptome in human liver disease.* Oncotarget, 2020. **11**(23): p. 2216-2232.

71. Lee, I., et al., *The Importance of Standardization on Analyzing Circulating RNA.* Molecular diagnosis & therapy, 2017. **21**(3): p. 259-268.

72. Rio, D.C., et al., *Purification of RNA using TRIzol (TRI reagent).* Cold Spring Harbor Protocols, 2010. **2010**(6): p. pdb. prot5439.

73. Chomczynski, P. and N. Sacchi, *Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.* Analytical Biochemistry, 1987. **162**(1): p. 156-159.

74. Tan, S.C. and B.C. Yiap, *DNA, RNA, and protein extraction: the past and the present.* Journal of biomedicine & biotechnology, 2009. **2009**: p. 574398-574398.

75. Berensmeier, S., *Magnetic particles for the separation and purification of nucleic acids.* Applied microbiology and biotechnology, 2006. **73**(3): p. 495-504.

76. Esser, K.-H., W.H. Marx, and T. Lisowsky, *maxXbond: first regeneration system for DNA binding silica matrices.* Nature methods, 2006. **3**(1): p. i-ii.

77. Bernard Lam, M.E., Song-Song Geng, Pan Robers, Nezer Rghei, and Yousef Haj-Ahmad. *Silicon Carbide as a Novel RNA Affinity Medium with Improved Sensitivity and Size Diversity*. [cited 2021 March 20]; Available from: https://norgenbiotek.com/sites/default/files/resources/Poster-13-Silicon-Carbide-as-a-Novel-RNA-Affinity-Medium-with-Improved-Sensitivity-and-Size-Diversity.pdf.

78. Wright, K., et al., *Comparison of methods for miRNA isolation and quantification from ovine plasma.* Scientific Reports, 2020. **10**(1): p. 825.

79. Srinivasan, S., et al., *Small RNA Sequencing across Diverse Biofluids Identifies Optimal Methods for exRNA Isolation.* Cell, 2019. **177**(2): p. 446-462.e16.

80. Fleige, S. and M.W. Pfaffl, *RNA integrity and the effect on the real-time qRT-PCR performance.* Molecular Aspects of Medicine, 2006. **27**(2): p. 126-139.

81. Gallagher, S.R., *Quantitation of DNA and RNA with Absorption and Fluorescence Spectroscopy.* Current Protocols in Human Genetics, 1994. **00**(1): p. A.3D.1-A.3D.8.

82. Wilfinger, W.W., K. Mackey, and P. Chomczynski, *Effect of pH and Ionic Strength on the Spectrophotometric Assessment of Nucleic Acid Purity.* BioTechniques, 1997. **22**(3): p. 474-481.

83. *Quantitating RNA.* [cited 2021 March 20]; Available from: https://www.thermofisher.com/us/en/home/references/ambion-tech-support/rna-isolation/tech-notes/quantitating-rna.html.

84. Glasel, J., *Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios.* Biotechniques, 1995. **18**(1): p. 62-63.

85. Manchester, K.L., *Use of UV methods for measurement of protein and nucleic acid concentrations.* Biotechniques, 1996. **20**(6): p. 968-970.

86. Singer, V.L., et al., *Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation.* Analytical biochemistry, 1997. **249**(2): p. 228-238.

87. Masotti, A. and T. Preckel, *Analysis of small RNAs with the Agilent 2100 Bioanalyzer.* Nature Methods, 2006. **3**(8): p. 658-658.

88. Gallagher, S.R. and P.R. Desjardins, *Quantitation of DNA and RNA with Absorption and Fluorescence Spectroscopy.* Current Protocols in Protein Science, 2008. **52**(1): p. A.4K.1-A.4K.21.

89. Shanker, S., et al., *Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA.* Journal of biomolecular techniques: JBT, 2015. **26**(1): p. 4.

90. Lightfoot, S. *Quantitation comparison of total RNA using the Agilent 2100 bioanalyzer, ribogreen analysis and UV spectrometry*. 2016 [cited 2021 April]; Available from: https://www.agilent.com/cs/library/applications/5988-7650EN.pdf.

91. Saiki, R.K., et al., *Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia.* Science, 1985. **230**(4732): p. 1350-4.

92. Kaunitz, J.D., *The Discovery of PCR: ProCuRement of Divine Power*. Digestive diseases and sciences, 2015. **60**(8): p. 2230-2231.

93. Liu, W. and D.A. Saint, *Validation of a quantitative method for real time PCR kinetics.* Biochemical and biophysical research communications, 2002. **294**(2): p. 347-353.

94. Kroneis, T., et al., *Global preamplification simplifies targeted mRNA quantification.* Scientific Reports, 2017. **7**(1): p. 45219.

95. *PCR Cycling Parameters–Six Key Considerations for Success*. [cited 2021 June 3]; Available from: https://www.thermofisher.com/us/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/pcr-education/pcr-reagents-enzymes/pcr-cycling-considerations.html.

96. Carter, N.P., et al., *Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer.* Genomics, 1992. **13**(3): p. 718-725.

97. Arya, M., et al., *Basic principles of real-time quantitative PCR.* Expert review of molecular diagnostics, 2005. **5**(2): p. 209-219.

98. Dragan, A., et al., *SYBR Green I: fluorescence properties and interaction with DNA.* Journal of fluorescence, 2012. **22**(4): p. 1189-1199.

99. Wang, Y., et al., *Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays.* BMC genomics, 2006. **7**(1): p. 1-16.

100. *Introduction to Quantitative PCR*. 2012 [cited 2021 April]; Available from: https://www.agilent.com/cs/library/brochures/Brochure_Guide%20to%20QPCR_IN70200C.pdf.

101. Rice, J., et al., *Assay Reproducibility in Clinical Studies of Plasma miRNA.* PLOS ONE, 2015. **10**(4): p. e0121948.

102. *Real-time PCR handbook*. 2012 [cited 2021 April]; Available from: https://www.gene-quantification.de/real-time-pcr-handbook-life-technologies-update-flr.pdf.

103. Everaert, C., et al., *Performance assessment of total RNA sequencing of human biofluids and extracellular vesicles.* Scientific reports, 2019. **9**(1): p. 1-16.

104. Alkhateeb, A., et al., *Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer.* Cancer informatics, 2019. **18**: p. 1176935119835522.

105. Makunin, J.M.I., *Non-coding RNA.* Hum Mol Genet, 2006. **15**(Spec No 1): p. R17-29.

106. Wapinski, O. and H.Y. Chang, *Long noncoding RNAs and human disease.* Trends in cell biology, 2011. **21**(6): p. 354-361.

107. Crowther, M.D., et al., *Genome-wide CRISPR–Cas9 screening reveals ubiquitous T cell cancer targeting via the monomorphic MHC class I-related protein MR1*. Nature Immunology, 2020. **21**(2): p. 178-185.

108. Jackson, H.W., et al., *The single-cell pathology landscape of breast cancer.* Nature, 2020. **578**(7796): p. 615-620.

109. Zhao, S., *Alternative splicing, RNA-seq and drug discovery*. Drug discovery today, 2019. **24**(6): p. 1258-1267.

110. Ye, C., et al., *DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery*. Nature communications, 2018. **9**(1): p. 1-9.

111. *SMARTer® Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian User Manual*. 2018 [cited 2021 April]; Available from: https://www.takarabio.com/documents/User%20Manual/SMARTer%20Stranded%20Total%20RNA/SMARTer%20Stranded%20Total%20RNA-Seq%20Kit%20v2%20-%20Pico%20Input%20Mammalian%20User%20Manual_050619.pdf.

112. Lin, X., et al., *A comparative analysis of RNA sequencing methods with ribosome RNA depletion for degraded and low-input total RNA from formalin-fixed and paraffin-embedded samples.* BMC genomics, 2019. **20**(1): p. 1-13.

113. Sarantopoulou, D., et al., *Comparative evaluation of RNA-Seq library preparation methods for strand-specificity and low input.* Scientific reports, 2019. **9**(1): p. 1-10.

114. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357.

115. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner*. Bioinformatics, 2013. **29**(1): p. 15-21.

116. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-1111.

117. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements.* Nature methods, 2015. **12**(4): p. 357-360.

118. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome biology, 2013. **14**(4): p. 1-13.

119. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 1-21.

120. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

121. Baranec, C., et al., *High-efficiency autonomous laser adaptive optics.* The Astrophysical Journal Letters, 2014. **790**(1): p. L8.

122. Li, X., et al., *Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies.* BMC Genomics, 2020. **21**(1): p. 75.

123. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biology, 2010. **11**(3): p. R25.

124. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome biology, 2010. **11**(3): p. 1-9.

125. Dillies, M.-A., et al., *A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.* Briefings in bioinformatics, 2013. **14**(6): p. 671-683.

126. Wagner, G.P., K. Kin, and V.J. Lynch, *Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.* Theory in biosciences, 2012. **131**(4): p. 281-285.

127. Usoskin, D., et al., *Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing.* Nature neuroscience, 2015. **18**(1): p. 145-153.

128. Abbas-Aghababazadeh, F., Q. Li, and B.L. Fridley, *Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing.* PLOS ONE, 2018. **13**(10): p. e0206312.

129. Lovén, J., et al., *Revisiting Global Gene Expression Analysis.* Cell, 2012. **151**(3): p. 476-482.

130. Risso, D., et al., *Normalization of RNA-seq data using factor analysis of control genes or samples.* Nature Biotechnology, 2014. **32**(9): p. 896-902.

131. Müller, J.B., et al., *Circulating biomarkers in patients with glioblastoma.* British journal of cancer, 2019.

132. Marcoux, G., et al., *Revealing the diversity of extracellular vesicles using high-dimensional flow cytometry analyses.* Scientific Reports, 2016. **6**(1): p. 35928.

133. Maas, S.L.N., X.O. Breakefield, and A.M. Weaver, *Extracellular Vesicles: Unique Intercellular Delivery Vehicles.* Trends in cell biology, 2017. **27**(3): p. 172-188.

134. Willms, E., et al., *Extracellular Vesicle Heterogeneity: Subpopulations, Isolation Techniques, and Diverse Functions in Cancer Progression.* Frontiers in immunology, 2018. **9**: p. 738-738.

135. Raposo, G. and W. Stoorvogel, *Extracellular vesicles: Exosomes, microvesicles, and friends.* The Journal of Cell Biology, 2013. **200**(4): p. 373-383.

136. McAndrews, K.M. and R. Kalluri, *Mechanisms associated with biogenesis of exosomes in cancer.* Molecular cancer, 2019. **18**(1): p. 52-52.

137. György, B., et al., *Membrane vesicles, current state-of-the-art: emerging role of extracellular vesicles.* Cellular and molecular life sciences : CMLS, 2011. **68**(16): p. 2667-2688.

138. Murillo, O.D., et al., *ExRNA atlas analysis reveals distinct extracellular RNA cargo types and their carriers present across human biofluids.* Cell, 2019. **177**(2): p. 463-477. e15.

139. Hinger, S.A., et al., *Diverse long RNAs are differentially sorted into extracellular vesicles secreted by colorectal cancer cells.* Cell reports, 2018. **25**(3): p. 715-725. e4.

140. Pös, O., et al., *Circulating cell-free nucleic acids: characteristics and applications.* European journal of human genetics : EJHG, 2018. **26**(7): p. 937-945.

141. Kishikawa, T., et al., *Circulating RNAs as new biomarkers for detecting pancreatic cancer.* World journal of gastroenterology, 2015. **21**(28): p. 8527-8540.

142. Ngo, T.T., et al., *Noninvasive blood tests for fetal development predict gestational age and preterm delivery.* Science, 2018. **360**(6393): p. 1133-1136.

143. Lin, J., et al., *Exosomes: novel biomarkers for clinical diagnosis.* The scientific world journal, 2015. **2015**.

144. Witwer, K.W., et al., *Standardization of sample collection, isolation and analysis methods in extracellular vesicle research.* Journal of extracellular vesicles, 2013. **2**(1): p. 20360.

145. Mitchell, A.J., et al., *Platelets confound the measurement of extracellular miRNA in archived plasma.* Scientific reports, 2016. **6**: p. 32651-32651.

146. Cheng, H.H., et al., *Plasma processing conditions substantially influence circulating microRNA biomarker levels.* PloS one, 2013. **8**(6): p. e64795-e64795.

147. Heijnen, H.F., et al., *Activated Platelets Release Two Types of Membrane Vesicles: Microvesicles by Surface Shedding and Exosomes Derived From Exocytosis of Multivesicular Bodies and-Granules.* Blood, The Journal of the American Society of Hematology, 1999. **94**(11): p. 3791-3799.

148. Menck, K., et al., *Isolation and Characterization of Microvesicles from Peripheral Blood.* Journal of visualized experiments : JoVE, 2017(119): p. 55057.

149. Brisson, A.R., et al., *Extracellular vesicles from activated platelets: a semiquantitative cryo-electron microscopy and immuno-gold labeling study.* Platelets, 2017. **28**(3): p. 263-271.

150. Menck, K., et al., *Characterisation of tumour-derived microvesicles in cancer patients' blood and correlation with clinical outcome.* Journal of Extracellular Vesicles, 2017. **6**(1): p. 1340745.

151. Aatonen, M.T., et al., *Isolation and characterization of platelet-derived extracellular vesicles.* Journal of extracellular vesicles, 2014. **3**: p. 10.3402/jev.v3.24692.

152. Théry, C., et al., *Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines.* Journal of Extracellular Vesicles, 2018. **7**(1): p. 1535750.

153. Hargett, L.A. and N.N. Bauer, *On the origin of microparticles: From "platelet dust" to mediators of intercellular communication.* Pulmonary circulation, 2013. **3**(2): p. 329-340.

154. Morgan, T.K., *Cell- and size-specific analysis of placental extracellular vesicles in maternal plasma and pre-eclampsia.* Translational Research, 2018. **201**: p. 40-48.

155. Robert, S., et al., *Standardization of platelet-derived microparticle counting using calibrated beads and a Cytomics FC500 routine flow cytometer: a first step towards multicenter studies?* Journal of Thrombosis and Haemostasis, 2009. **7**(1): p. 190-197.

156. Poncelet, P., et al., *Standardized counting of circulating platelet microparticles using currently available flow cytometers and scatter-based triggering: Forward or side scatter?* Cytometry Part A, 2016. **89**(2): p. 148-158.

157. Jayachandran, M., et al., *Methodology for isolation, identification and characterization of microvesicles in peripheral blood.* Journal of immunological methods, 2012. **375**(1-2): p. 207-214.

158. Van Der Vlist, E.J., et al., *Fluorescent labeling of nano-sized vesicles released by cells and subsequent quantitative and qualitative analysis by high-resolution flow cytometry.* Nature protocols, 2012. **7**(7): p. 1311-1326.

159. van der Pol, E., et al., *Standardization of extracellular vesicle measurements by flow cytometry through vesicle diameter approximation.* J Thromb Haemost, 2018. **16**(6): p. 1236-1245.

160. van Manen, H.-J., et al., *Refractive Index Sensing of Green Fluorescent Proteins in Living Cells Using Fluorescence Lifetime Imaging Microscopy.* Biophysical Journal, 2008. **94**(8): p. L67-L69.

161. Mitchell, P.S., et al., *Circulating microRNAs as stable blood-based markers for cancer detection.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(30): p. 10513-10518.

162. Souza, M.F.D., et al., *Circulating mRNA signature as a marker for high-risk prostate cancer.* Carcinogenesis, 2019.

163. Xue, V.W., et al., *Non-invasive Potential Circulating mRNA Markers for Colorectal Adenoma Using Targeted Sequencing.* Scientific Reports, 2019. **9**(1): p. 12943.

164. Andaloussi, S.E., et al., *Extracellular vesicles: biology and emerging therapeutic opportunities.* Nature reviews Drug discovery, 2013. **12**(5): p. 347-357.

165. Zaporozhchenko, I.A., et al., *The potential of circulating cell-free RNA as a cancer biomarker: challenges and opportunities.* Expert review of molecular diagnostics, 2018. **18**(2): p. 133-145.

166. Coumans, F.A., et al., *Methodological guidelines to study extracellular vesicles.* Circulation research, 2017. **120**(10): p. 1632-1648.

167. Welsh, J.A., et al., *MIFlowCyt-EV: a framework for standardized reporting of extracellular vesicle flow cytometry experiments.* Journal of Extracellular Vesicles, 2020. **9**(1): p. 1713526.

168. Bode, A.P., et al., *Vesiculation of platelets during in vitro aging.* Blood, 1991. **77**(4): p. 887-95.

169. Thompson, C.B., et al., *Size dependent platelet subpopulations: relationship of platelet volume to ultrastructure, enzymatic activity, and function.* British Journal of Haematology, 1982. **50**(3): p. 509-519.

170. Sheridan, C., *Investors keep the faith in cancer liquid biopsies.* Nature biotechnology, 2019. **37**(9): p. 972.

171. Loeb, S., et al., *Complications after prostate biopsy: data from SEER-Medicare.* The Journal of urology, 2011. **186**(5): p. 1830-1834.

172. Berger-Richardson, D. and C.J. Swallow, *Needle tract seeding after percutaneous biopsy of sarcoma: risk/benefit considerations.* Cancer, 2017. **123**(4): p. 560-567.

173. Brenner, D.J., *Radiation risks potentially associated with low-dose CT screening of adult smokers for lung cancer.* Radiology, 2004. **231**(2): p. 440-445.

174. Hendrick, R.E., *Radiation doses and cancer risks from breast imaging studies.* Radiology, 2010. **257**(1): p. 246-253.

175. Etzioni, R., et al., *Early detection: The case for early detection.* Nature Reviews Cancer, 2003. **3**(4): p. 243.

176. Crowley, E., et al., *Liquid biopsy: monitoring cancer-genetics in the blood.* Nature reviews Clinical oncology, 2013. **10**(8): p. 472.

177. Zemmour, H., et al., *Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA.* Nature communications, 2018. **9**(1): p. 1443.

178. Butler, T.M., et al., *Exome sequencing of cell-free DNA from metastatic cancer patients identifies clinically actionable mutations distinct from primary disease.* PloS one, 2015. **10**(8): p. e0136407.

179. Newman, A.M., et al., *An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage.* Nature medicine, 2014. **20**(5): p. 548.

180. Arroyo, J.D., et al., *Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma.* Proceedings of the National Academy of Sciences, 2011. **108**(12): p. 5003-5008.

181. Williams, Z., et al., *Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations.* Proceedings of the National Academy of Sciences, 2013. **110**(11): p. 4255-4260.

182. Tsang, J.C., et al., *Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics.* Proceedings of the National Academy of Sciences, 2017. **114**(37): p. E7786-E7795.

183. Koh, W., et al., *Noninvasive in vivo monitoring of tissue-specific global gene expression in humans.* Proceedings of the National Academy of Sciences, 2014. **111**(20): p. 7361-7366.

184. Yang, K.S., et al., *Multiparametric plasma EV profiling facilitates diagnosis of pancreatic malignancy.* Science translational medicine, 2017. **9**(391): p. eaal3226.

185. Lázaro-Ibáñez, E., et al., *Different gDNA content in the subpopulations of prostate cancer extracellular vesicles: apoptotic bodies, microvesicles, and exosomes.* The Prostate, 2014. **74**(14): p. 1379-1390.

186. Scheer, F.A., et al., *Impact of the human circadian system, exercise, and their interaction on cardiovascular function.* Proceedings of the National Academy of Sciences, 2010. **107**(47): p. 20541-20546.

187. Lange, T., S. Dimitrov, and J. Born, *Effects of sleep and circadian rhythm on the human immune system.* Annals of the New York Academy of Sciences, 2010. **1193**(1): p. 48-59.

188. Vollmers, C., et al., *Time of feeding and the intrinsic circadian clock drive rhythms in hepatic gene expression.* Proceedings of the National Academy of Sciences, 2009. **106**(50): p. 21453-21458.

189. Buhr, E.D. and J.S. Takahashi, *Molecular components of the Mammalian circadian clock*, in *Circadian clocks*. 2013, Springer. p. 3-27.

190. Floris, I., et al., *MiRNA analysis by quantitative PCR in preterm human breast milk reveals daily fluctuations of hsa-miR-16-5p.* PLoS One, 2015. **10**(10): p. e0140488.

191. Heegaard, N.H., et al., *Diurnal variations of human circulating cell-free micro-RNA.* PloS one, 2016. **11**(8): p. e0160577.

192. Hindson, B.J., et al., *High-throughput droplet digital PCR system for absolute quantitation of DNA copy number.* Analytical chemistry, 2011. **83**(22): p. 8604-8610.

193. Taylor, S.C., G. Laperriere, and H. Germain, *Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data.* Scientific reports, 2017. **7**(1): p. 2409.

194. Olmedillas-López, S., M. García-Arranz, and D. García-Olmo, *Current and emerging applications of droplet digital PCR in oncology.* Molecular diagnosis & therapy, 2017. **21**(5): p. 493-510.

195. Welsh, J.A., et al., *MIFlowCyt-EV: a framework for standardized reporting of extracellular vesicle flow cytometry experiments.* Journal of Extracellular Vesicles, 2020. **9**(1): p. 1713526.

196. Aatonen, M.T., et al., *Isolation and characterization of platelet-derived extracellular vesicles.* Journal of extracellular vesicles, 2014. **3**(1): p. 24692.

197. Devonshire, A.S., et al., *Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification.* Analytical and bioanalytical chemistry, 2014. **406**(26): p. 6499-6512.

198. Hothorn, T., et al., *A lego system for conditional inference.* The American Statistician, 2006. **60**(3): p. 257-263.

199. Trivedi, D.K., et al., *Discovery of volatile biomarkers of Parkinson's disease from sebum.* ACS Central Science, 2019. **5**(4): p. 599-606.

200. van Ginkel, J.H., et al., *Preanalytical blood sample workup for cell-free DNA analysis using Droplet Digital PCR for future molecular cancer diagnostics.* Cancer medicine, 2017. **6**(10): p. 2297-2307.

201. Welsh, J.A., et al., *Extracellular vesicle flow cytometry analysis and standardization.* Frontiers in cell and developmental biology, 2017. **5**: p. 78.

202. Max, K.E., et al., *Human plasma and serum extracellular small RNA reference profiles and their clinical utility.* Proceedings of the National Academy of Sciences, 2018. **115**(23): p. E5334-E5343.

203. Madsen, A.T., et al., *Day-to-day and within-day biological variation of cell-free DNA.* EBioMedicine, 2019. **49**: p. 284-290.

204. Zhong, X.Y., et al., *Fluctuation of maternal and fetal free extracellular circulatory DNA in maternal plasma.* Obstetrics & Gynecology, 2000. **96**(6): p. 991-996.

205. Chen, M., et al., *Utility of Circulating Cell-Free RNA Analysis for the Characterization of Global Transcriptome Profiles of Multiple Myeloma Patients.* Cancers, 2019. **11**(6): p. 887.

206. Li, Y., et al., *Serum circulating human mRNA profiling and its utility for oral cancer detection.* J Clin Oncol, 2006. **24**(11): p. 1754-1760.

207. Dheda, K., et al., *Validation of housekeeping genes for normalizing RNA expression in real-time PCR.* Biotechniques, 2004. **37**(1): p. 112-119.

208. De Jonge, H.J., et al., *Evidence based selection of housekeeping genes.* PloS one, 2007. **2**(9): p. e898.

209. Yang, Q., et al., *Evaluation and validation of the suitable control genes for quantitative PCR studies in plasma DNA for non-invasive prenatal diagnosis.* International journal of molecular medicine, 2014. **34**(6): p. 1681-1687.

210. Morgan, T.K., *Cell-and size-specific analysis of placental extracellular vesicles in maternal plasma and pre-eclampsia.* Translational Research, 2018. **201**: p. 40-48.

211. Danielson, K.M., et al., *Diurnal variations of circulating extracellular vesicles measured by nano flow cytometry.* PloS one, 2016. **11**(1): p. e0144678.

212. Doyle, L.M. and M.Z. Wang, *Overview of extracellular vesicles, their origin, composition, purpose, and methods for exosome isolation and analysis.* Cells, 2019. **8**(7): p. 727.

213. Herbst, R.S., D. Morgensztern, and C. Boshoff, *The biology and management of non-small cell lung cancer.* Nature, 2018. **553**: p. 446.

214. Lennon, F.E., et al., *Lung cancer-a fractal viewpoint.* Nat Rev Clin Oncol, 2015. **12**(11): p. 664-75.

215. Howlader N, N.A., Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA *SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD.*

216. Kyle, R.A. and S.V. Rajkumar, *Management of monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM).* Oncology (Williston Park), 2011. **25**(7): p. 578-86.

217. Dhodapkar, M.V., *MGUS to myeloma: a mysterious gammopathy of underexplored significance.* Blood, 2016. **128**(23): p. 2599.

218. Llovet, J.M., et al., *Hepatocellular carcinoma.* Nat Rev Dis Primers, 2016. **2**: p. 16018.

219. Fateen, W. and S.D. Ryder, *Screening for hepatocellular carcinoma: patient selection and perspectives.* J Hepatocell Carcinoma, 2017. **4**: p. 71-79.

220. Starr, S.P. and D. Raines, *Cirrhosis: diagnosis, management, and prevention.* Am Fam Physician, 2011. **84**(12): p. 1353-9.

221. Laursen, L., *A preventable cancer.* Nature, 2014. **516**: p. S2.

222. Goh, G.B., P.E. Chang, and C.K. Tan, *Changing epidemiology of hepatocellular carcinoma in Asia.* Best Pract Res Clin Gastroenterol, 2015. **29**(6): p. 919-28.

223. Wong, V.W., et al., *Clinical scoring system to predict hepatocellular carcinoma in chronic hepatitis B carriers.* J Clin Oncol, 2010. **28**(10): p. 1660-5.

224. Yang, H.I., et al., *Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): development and validation of a predictive score.* Lancet Oncol, 2011. **12**(6): p. 568-74.

225. Bai, Y. and H. Zhao, *Liquid biopsy in tumors: opportunities and challenges.* Ann Transl Med, 2018. **6**(Suppl 1): p. S89.

226. Palmirotta, R., et al., *Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology.* Ther Adv Med Oncol, 2018. **10**: p. 1758835918794630.

227. Marrugo-Ramírez, J., M. Mir, and J. Samitier, *Blood-Based Cancer Biomarkers in Liquid Biopsy: A Promising Non-Invasive Alternative to Tissue Biopsy.* Int J Mol Sci, 2018. **19**(10).

228. Esposito, A., et al., *Liquid biopsies for solid tumors: Understanding tumor heterogeneity and real time monitoring of early resistance to targeted therapies.* Pharmacol Ther, 2016. **157**: p. 120-4.

229. Sundling, K.E. and A.C. Lowe, *Circulating Tumor Cells: Overview and Opportunities in Cytology.* Adv Anat Pathol, 2019. **26**(1): p. 56-63.

230. Millner, L.M., M.W. Linder, and R. Valdes, Jr., *Circulating tumor cells: a review of present methods and the need to identify heterogeneous phenotypes.* Ann Clin Lab Sci, 2013. **43**(3): p. 295-304.

231. Thiele, J.A., et al., *Circulating Tumor Cells: Fluid Surrogates of Solid Tumors.* Annual Review of Pathology: Mechanisms of Disease, 2017. **12**(1): p. 419-447.

232. Liu, Y. and X. Cao, *The origin and function of tumor-associated macrophages.* Cellular And Molecular Immunology, 2014. **12**: p. 1.

233. Adams, D.L., et al., *Circulating giant macrophages as a potential biomarker of solid tumors.* Proceedings of the National Academy of Sciences, 2014. **111**(9): p. 3514.

234. Gast, C.E., et al., *Cell fusion potentiates tumor heterogeneity and reveals circulating hybrid cells that correlate with stage and survival.* Science Advances, 2018. **4**(9): p. eaat7828.

235. Newman, A.M., et al., *An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage.* Nat Med, 2014. **20**(5): p. 548-54.

236. Corcoran, R.B. and B.A. Chabner, *Application of Cell-free DNA Analysis to Cancer Treatment.* N Engl J Med, 2018. **379**(18): p. 1754-1765.

237. Abbosh, C., et al., *Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution.* Nature, 2017. **545**(7655): p. 446-451.

238. Best, M.G., et al., *RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics.* Cancer Cell, 2015. **28**(5): p. 666-676.

239. Best, M.G., P. Wesseling, and T. Wurdinger, *Tumor-Educated Platelets as a Noninvasive Biomarker Source for Cancer Detection and Progression Monitoring.* Cancer Res, 2018. **78**(13): p. 3407-3412.

240. In, S.G.J.G. t Veld, and T. Wurdinger, *Tumor-educated platelets.* Blood, 2019: p. blood-2018-12-852830.

241. Cohen, J.D., et al., *Detection and localization of surgically resectable cancers with a multi-analyte blood test.* Science, 2018. **359**(6378): p. 926.

242. Abbosh, C., N.J. Birkbak, and C. Swanton, *Early stage NSCLC - challenges to implementing ctDNA-based screening and MRD detection.* Nat Rev Clin Oncol, 2018. **15**(9): p. 577-586.

243. Haque, I.S. and O. Elemento, *Challenges in Using ctDNA to Achieve Early Detection of Cancer.* bioRxiv, 2017: p. 237578.

244. Salta, S., et al., *A DNA Methylation-Based Test for Breast Cancer Detection in Circulating Cell-Free DNA.* J Clin Med, 2018. **7**(11).

245. Xu, R.-h., et al., *Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma.* Nature Materials, 2017. **16**: p. 1155.

246. Song, C.-X., et al., *5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages.* Cell Research, 2017. **27**: p. 1231.

247. Shen, S.Y., et al., *Sensitive tumour detection and classification using plasma cell-free DNA methylomes.* Nature, 2018. **563**(7732): p. 579-583.

248. Moss, J., et al., *Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease.* Nature Communications, 2018. **9**(1): p. 5068.

249. Cristiano, S., et al., *Genome-wide cell-free DNA fragmentation in patients with cancer.* Nature, 2019. **570**(7761): p. 385-389.

250. Gemmell, C.H., *Activation of platelets by in vitro whole blood contact with materials: increases in microparticle, procoagulant activity, and soluble P-selectin blood levels.* J Biomater Sci Polym Ed, 2001. **12**(8): p. 933-43.

251. Heitzer, E., et al., *Current and future perspectives of liquid biopsies in genomics-driven oncology.* Nature Reviews Genetics, 2019. **20**(2): p. 71-88.

252. Wan, J.C.M., et al., *Liquid biopsies come of age: towards implementation of circulating tumour DNA.* Nature Reviews Cancer, 2017. **17**: p. 223.

253. Koh, W., et al., *Noninvasive in vivo monitoring of tissue-specific global gene expression in humans.* Proceedings of the National Academy of Sciences, 2014: p. 201405528.

254. Pan, W., et al., *Simultaneously Monitoring Immune Response and Microbial Infections During Pregnancy through Plasma cfRNA Sequencing.* Clinical Chemistry, 2016: p. clinchem.2017.273888.

255. Ngo, T.T.M., et al., *Noninvasive blood tests for fetal development predict gestational age and preterm delivery.* Science, 2018. **360**(6393): p. 1133.

256. Joshi NA, F.J., *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.* Available at https://github.com/najoshi/sickle, 2011.

257. Leggett, R.M., et al., *Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics.* Frontiers in genetics, 2013. **4**: p. 288-288.

258. Andrews, S., *FastQC: a quality control tool for high throughput sequence data.* 2010.

259. Wang, L., S. Wang, and W. Li, *RSeQC: quality control of RNA-seq experiments.* Bioinformatics, 2012. **28**(16): p. 2184-5.

260. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11.10.1-33.

261. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

262. Wang, L., S. Wang, and W. Li, *RSeQC: quality control of RNA-seq experiments.* Bioinformatics, 2012. **28**(16): p. 2184-2185.

263. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

264. Kuhn, M., *Building Predictive Models in R Using the caret Package.* Journal of Statistical Software, 2008. **28**(5).

265. Alexa A, R.J., *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.36.0., 2019.

266. Ripley, W.N.V.a.B.D., *Modern Applied Statistics with S.* 2002.

267. Wiener, A.L.a.M., *Classification and Regression by randomForest.* R News, 2002. **2**(3): p. 18-22.

268. *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.

269. Sardar, H.S. and S.P. Gilbert, *Microtubule capture by mitotic kinesin centromere protein E (CENP-E).* J Biol Chem, 2012. **287**(30): p. 24894-904.

270. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.

271. Fry, A.M., *The Nek2 protein kinase: a novel regulator of centrosome structure.* Oncogene, 2002. **21**(40): p. 6184-6194.

272. Mills, C.A., et al., *Nucleolar and spindle-associated protein 1 (NUSAP1) interacts with a SUMO E3 ligase complex during chromosome segregation.* J Biol Chem, 2017. **292**(42): p. 17178-17189.

273. Srivastava, R.A., N. Bhasin, and N. Srivastava, *Apolipoprotein E gene expression in various tissues of mouse and regulation by estrogen.* Biochem Mol Biol Int, 1996. **38**(1): p. 91-101.

274. Jia, Q., et al., *Association between complement C3 and prevalence of fatty liver disease in an adult population: a cross-sectional study from the Tianjin Chronic Low-Grade Systemic Inflammation and Health (TCLSIHealth) cohort study.* PLoS One, 2015. **10**(4): p. e0122026.

275. Zeng, D.W., et al., *Serum ceruloplasmin levels correlate negatively with liver fibrosis in males with chronic hepatitis B: a new noninvasive model for predicting liver fibrosis in HBV-related liver disease.* PLoS One, 2013. **8**(10): p. e77942.

276. Waterham, H.R., et al., *Mutations in the 3beta-hydroxysterol Delta24-reductase gene cause desmosterolosis, an autosomal recessive disorder of cholesterol biosynthesis.* Am J Hum Genet, 2001. **69**(4): p. 685-94.

277. Fort, A., et al., *A liver enhancer in the fibrinogen gene cluster.* Blood, 2011. **117**(1): p. 276-82.

278. Gram, J., et al., *Plasma histidine-rich glycoprotein and plasminogen in patients with liver disease.* Thromb Res, 1985. **39**(4): p. 411-7.

279. Goodman, Z.D., *Liver Biopsy Diagnosis of Cirrhosis*, in *Diagnostic Methods for Cirrhosis and Portal Hypertension*, A. Berzigotti and J. Bosch, Editors. 2018, Springer International Publishing: Cham. p. 17-31.

280. Funaki, N.O., et al., *Identification of carcinoembryonic antigen mRNA in circulating peripheral blood of pancreatic carcinoma and gastric carcinoma patients.* Life Sci, 1996. **59**(25-26): p. 2187-99.

281. Kishikawa, T., et al., *Circulating RNAs as new biomarkers for detecting pancreatic cancer.* World J Gastroenterol, 2015. **21**(28): p. 8527-40.

282. Kopreski, M.S., et al., *Detection of tumor messenger RNA in the serum of patients with malignant melanoma.* Clin Cancer Res, 1999. **5**(8): p. 1961-5.

283. Lo, K.W., et al., *Analysis of cell-free Epstein-Barr virus associated RNA in the plasma of patients with nasopharyngeal carcinoma.* Clin Chem, 1999. **45**(8 Pt 1): p. 1292-4.

284. Yu, S., et al., *Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma.* Gut, 2020. **69**(3): p. 540-550.

285. Enderle, D., et al., *Characterization of RNA from exosomes and other extracellular vesicles isolated by a novel spin column-based method.* PloS one, 2015. **10**(8): p. e0136133.

286. Mantel, P.-Y., et al., *Infected erythrocyte-derived extracellular vesicles alter vascular function via regulatory Ago2-miRNA complexes in malaria.* Nature communications, 2016. **7**(1): p. 1-15.

287. Das, S., et al., *The extracellular RNA communication consortium: establishing foundational knowledge and technologies for extracellular RNA research.* Cell, 2019. **177**(2): p. 231-242.

288. Zhang, Q., et al., *Transfer of Functional Cargo in Exomeres.* Cell Reports, 2019. **27**(3): p. 940-954.e6.

289. Lässer, C., *Mapping extracellular rna sheds lights on distinct carriers.* Cell, 2019. **177**(2): p. 228-230.

290. Yang, Z., et al., *A Multianalyte Panel Consisting of Extracellular Vesicle miRNAs and mRNAs, cfDNA, and CA19-9 Shows Utility for Diagnosis and Staging of Pancreatic Ductal Adenocarcinoma.* Clinical Cancer Research, 2020. **26**(13): p. 3248-3258.

291. Melo, S.A., et al., *Cancer exosomes perform cell-independent microRNA biogenesis and promote tumorigenesis.* Cancer cell, 2014. **26**(5): p. 707-721.

292. Geekiyanage, H., et al., *Extracellular microRNAs in human circulation are associated with miRISC complexes that are accessible to anti-AGO2 antibody and can bind target mimic oligonucleotides.* Proceedings of the National Academy of Sciences, 2020. **117**(39): p. 24213-24223.

293. Temoche-Diaz, M.M., et al., *Distinct mechanisms of microRNA sorting into cancer cell-derived extracellular vesicle subtypes.* Elife, 2019. **8**.

294. Hutvagner, G. and M.J. Simard, *Argonaute proteins: key players in RNA silencing.* Nat Rev Mol Cell Biol, 2008. **9**(1): p. 22-32.

295. Hulstaert, E., et al., *Charting Extracellular Transcriptomes in The Human Biofluid RNA Atlas.* Cell Reports, 2020. **33**(13): p. 108552.

296. Vickers, K.C., et al., *MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins.* Nat Cell Biol, 2011. **13**(4): p. 423-33.

297. Fong, M.Y., et al., *Breast-cancer-secreted miR-122 reprograms glucose metabolism in premetastatic niche to promote metastasis.* Nat Cell Biol, 2015. **17**(2): p. 183-94.

298. Aizarani, N., et al., *A human liver cell atlas reveals heterogeneity and epithelial progenitors.* Nature, 2019. **572**(7768): p. 199-204.

299. Hsiao, L.L., et al., *A compendium of gene expression in normal human tissues.* Physiol Genomics, 2001. **7**(2): p. 97-104.

300. Acevedo, L.G., et al., *Analysis of the mechanisms mediating tumor-specific changes in gene expression in human liver tumors.* Cancer Res, 2008. **68**(8): p. 2641-51.

301. Hoshida, Y., et al., *Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma.* Cancer Res, 2009. **69**(18): p. 7385-92.

302. Lee, J.S., et al., *Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling.* Hepatology, 2004. **40**(3): p. 667-76.

303. Director's Challenge Consortium for the Molecular Classification of Lung, A., et al., *Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.* Nature medicine, 2008. **14**(8): p. 822-827.

304. Sarrió, D., et al., *Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype.* Cancer Res, 2008. **68**(4): p. 989-97.

305. Risso, D., et al., *Normalization of RNA-seq data using factor analysis of control genes or samples.* Nat Biotechnol, 2014. **32**(9): p. 896-902.

306. Tickner, J.A., et al., *Functions and therapeutic roles of exosomes in cancer.* Frontiers in oncology, 2014. **4**: p. 127-127.

307. Zhang, X., et al., *Exosomes in cancer: small particle, big player.* Journal of hematology & oncology, 2015. **8**: p. 83-83.