Cell

exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids

Oscar D. Murillo,[1,31] William Thistlethwaite,[1,31] Joel Rozowsky,[2] Sai Lakshmi Subramanian,[1] Rocco Lucero,[1] Neethu Shah,[1] Andrew R. Jackson,[1] Srimeenakshi Srinivasan,[3] Allen Chung,[4,5] Clara D. Laurent,[3] Robert R. Kitchen,[6] Timur Galeev,[2] Jonathan Warrell,[2] James A. Diao,[2,7] Joshua A. Welsh,[8] Kristina Hanspers,[9] Anders Riutta,[9] Sebastian Burgstaller-Muehlbacher,[10] Ravi V. Shah,[11] Ashish Yeri,[11] Lisa M. Jenkins,[12] Mehmet E. Ahsen,[13] Carlos Cordon-Cardo,[14] Navneet Dogra,[13,15] Stacey M. Gifford,[15] Joshua T. Smith,[15] Gustavo Stolovitzky,[13,15] Ashutosh K. Tewari,[16] Benjamin H. Wunsch,[15] Kamlesh K. Yadav,[16,17] Kirsty M. Danielson,[11]

(Author list continued on next page)

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[3]Department of Obstetrics, Gynecology, and Reproductive Sciences and Sanford Consortium for Regenerative Medicine, University of California, San Diego, La Jolla, CA 92037, USA
[4]Department of Surgery, University of California, San Francisco, San Francisco, CA 94143, USA
[5]Surgical Service, San Francisco Veterans Affairs Medical Center, San Francisco, CA 94121, USA
[6]Exosome Diagnostics, Inc., Waltham, MA 02451, USA
[7]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
[8]Translational Nanobiology Section, Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

(Affiliations continued on next page)

SUMMARY

To develop a map of cell-cell communication mediated by extracellular RNA (exRNA), the NIH Extracellular RNA Communication Consortium created the exRNA Atlas resource (https://exrna-atlas.org). The Atlas version 4P1 hosts 5,309 exRNA-seq and exRNA qPCR profiles from 19 studies and a suite of analysis and visualization tools. To analyze variation between profiles, we apply computational deconvolution. The analysis leads to a model with six exRNA cargo types (CT1, CT2, CT3A, CT3B, CT3C, CT4), each detectable in multiple biofluids (serum, plasma, CSF, saliva, urine). Five of the cargo types associate with known vesicular and non-vesicular (lipoprotein and ribonucleoprotein) exRNA carriers. To validate utility of this model, we re-analyze an exercise response study by deconvolution to identify physiologically relevant response pathways that were not detected previously. To enable wide application of this model, as part of the exRNA Atlas resource, we provide tools for deconvolution and analysis of user-provided case-control studies.

INTRODUCTION

The Extracellular RNA Communication Consortium (ERCC) (Ainsztein et al., 2015) aims to realize the potential of extracellular RNA (exRNA) as disease indicators and therapeutic molecules and to define the fundamental principles of their biogenesis, distribution, uptake, and function. In the context of this overall effort, the ERCC has developed the Extracellular RNA Atlas, a reference catalog of exRNAs present in human biofluids. The current version of the Atlas provides access to 5,309 exRNA-seq and exRNA qPCR sample profiles, primarily from cerebrospinal fluid (CSF), saliva, serum, plasma, and urine, collected across 19 different studies. A suite of web-accessible tools enables users to analyze exRNA-seq profiles from the Atlas, process and analyze their own exRNA-seq data, and contribute their data and analysis results to the Atlas, thus creating a virtuous cycle of knowledge creation.

One major obstacle encountered early on in the project was the large unexplained variability in exRNA profiles both within and across exRNA profiling studies. This variability diminishes the power of individual exRNA profiling studies and the utility of the Atlas as a source of exRNA reference profiles for detecting disease-associated perturbations. In an attempt to minimize experimental variability, all exRNA-seq profiles were quality-controlled and uniformly processed using the exceRpt pipeline (Rozowsky et al., 2019). However, the uniform processing through exceRpt failed to explain a large amount of residual sample-to-sample and between-study variability. We reasoned that much of the residual variability may be due to the combined effect of (1) known differences in the exRNA cargo profiles of vesicular (Lässer et al., 2017) and non-vesicular (Allen et al., 2018) exRNA carriers, and (2) technical or biological variation in carrier proportions between individual samples and across studies. To explore this hypothesis, we began with the results of previous studies that profiled different physically isolated carriers of

Justyna Filant,[18] Courtney Moeller,[3] Parham Nejad,[19] Anu Paul,[19] Bridget Simonson,[11] David K. Wong,[4,5] Xuan Zhang,[6] Leonora Balaj,[20] Roopali Gandhi,[19] Anil K. Sood,[18,21,22] Roger P. Alexander,[23] Liang Wang,[24] Chunlei Wu,[10] David T.W. Wong,[25] David J. Galas,[23] Kendall Van Keuren-Jensen,[26] Tushar Patel,[27] Jennifer C. Jones,[8] Saumya Das,[11] Kei-Hoi Cheung,[28] Alexander R. Pico,[9] Andrew I. Su,[10] Robert L. Raffai,[4,5] Louise C. Laurent,[3] Matthew E. Roth,[1] Mark B. Gerstein,[2,29,30] and Aleksandar Milosavljevic[1,32,*]

[9]Gladstone Institutes, San Francisco, CA 94158, USA
[10]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA
[11]Cardiovascular Research Center, Cardiology Division, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA
[12]Laboratory of Cell Biology, Center for Cancer Research, NIH, Bethesda, MD 20892, USA
[13]Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[14]Department of Pathology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[15]IBM T.J. Watson Research Center, IBM Research, Yorktown Heights, NY 10598, USA
[16]Department of Urology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[17]Sema4, Stamford, CT 06902, USA
[18]Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[19]Department of Neurology, Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[20]Department of Neurosurgery, Massachusetts General Hospital, Boston, MA 02114, USA
[21]Center for RNA Interference and Non-Coding RNAs, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[22]Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[23]Pacific Northwest Research Institute, Seattle, WA 98122, USA
[24]Department of Pathology and MCW Cancer Center, Medical College of Wisconsin, Milwaukee, WI 53226, USA
[25]School of Dentistry, University of California, Los Angeles, Los Angeles, CA 90095, USA
[26]Neurogenomics, The Translational Genomics Research Institute (TGen), Phoenix, AZ 85004, USA
[27]Department of Transplantation, Mayo Clinic, Jacksonville, FL 32224, USA
[28]Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT 06520, USA
[29]Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT 06520, USA
[30]Department of Computer Science, Yale University, New Haven, CT 06520, USA
[31]These authors contributed equally
[32]Lead Contact
*Correspondence: amilosav@bcm.edu
https://doi.org/10.1016/j.cell.2019.02.018

extracellular RNA, including high-density vesicles (HDVs) and low-density vesicles (LDVs) in the supernatant of the HMC-1 mast cell line (Lässer et al., 2017) and high-density lipoprotein (HDL) particles. We adapted a computational deconvolution method (Onuchic et al., 2016) to uncover the contributions of these known carrier-specific cargo types to each sample in the Atlas, as well as estimate new cargo types that are yet to be physically isolated and profiled. Computational deconvolution explained a large fraction of sample-to-sample and between-study variability in the exRNA Atlas profiles and suggested a model consisting of six cargo types, each detectable across studies and in at least two human biofluids. Additional separation experiments suggest association of the cargo types with specific vesicular and non-vesicular (RNP and lipoprotein) exRNA carriers. We show that these findings and Atlas resources facilitate interpretation of exRNA profiling studies by providing estimates of cargo type proportions in each sample and by tracing exRNA differences between cases and controls to specific cargo types and their associated carriers.

RESULTS

exRNA Atlas Resource

The exRNA Atlas resource is the data repository of the ERCC and integrates tools, web services, and pathway knowledge relevant for collaborative extracellular RNA research. The Atlas software is a free open source Genboree application supported by the document-oriented Genboree KnowledgeBase (GenboreeKB) back-end. Version 4P1 of the Atlas (https://exrna-atlas.org/exatv4p1) contains 2,270 exRNA-seq and 3,039 qPCR profiles from 19 different studies that cover 23 health conditions. Samples come predominantly from five biofluids (CSF, plasma, saliva, serum, urine) and have been collected primarily from total cell-free biofluid RNA using a variety of RNA isolation kits (Figure 1A) and processed using various sequencing library preparation kits (Table S1). The exRNA Atlas metadata follow the definitions established by the Metadata Working Group (MWG) of the ERCC utilizing Gene Ontology (GO) (Ashburner et al., 2000) to classify exRNA carriers (e.g., extracellular exosome [GO:0070062], extracellular vesicles [GO:1903561], and HDL-containing protein-lipid-RNA complex [GO:1990685]), relationships relevant to exRNA source, and new terms used to annotate samples (Cheung et al., 2016). Recently, the extracellular vesicle community published updated guidelines on the minimal information required for studies of extracellular vesicles (MISEV) (Théry et al., 2018); the ERCC will continue to update the exRNA Atlas resource accordingly, as MISEV guidelines evolve. A summary of exRNA Atlas samples, including biofluid type, methods of RNA isolation, percent passed quality control (QC), and reads mapped per RNA biotype

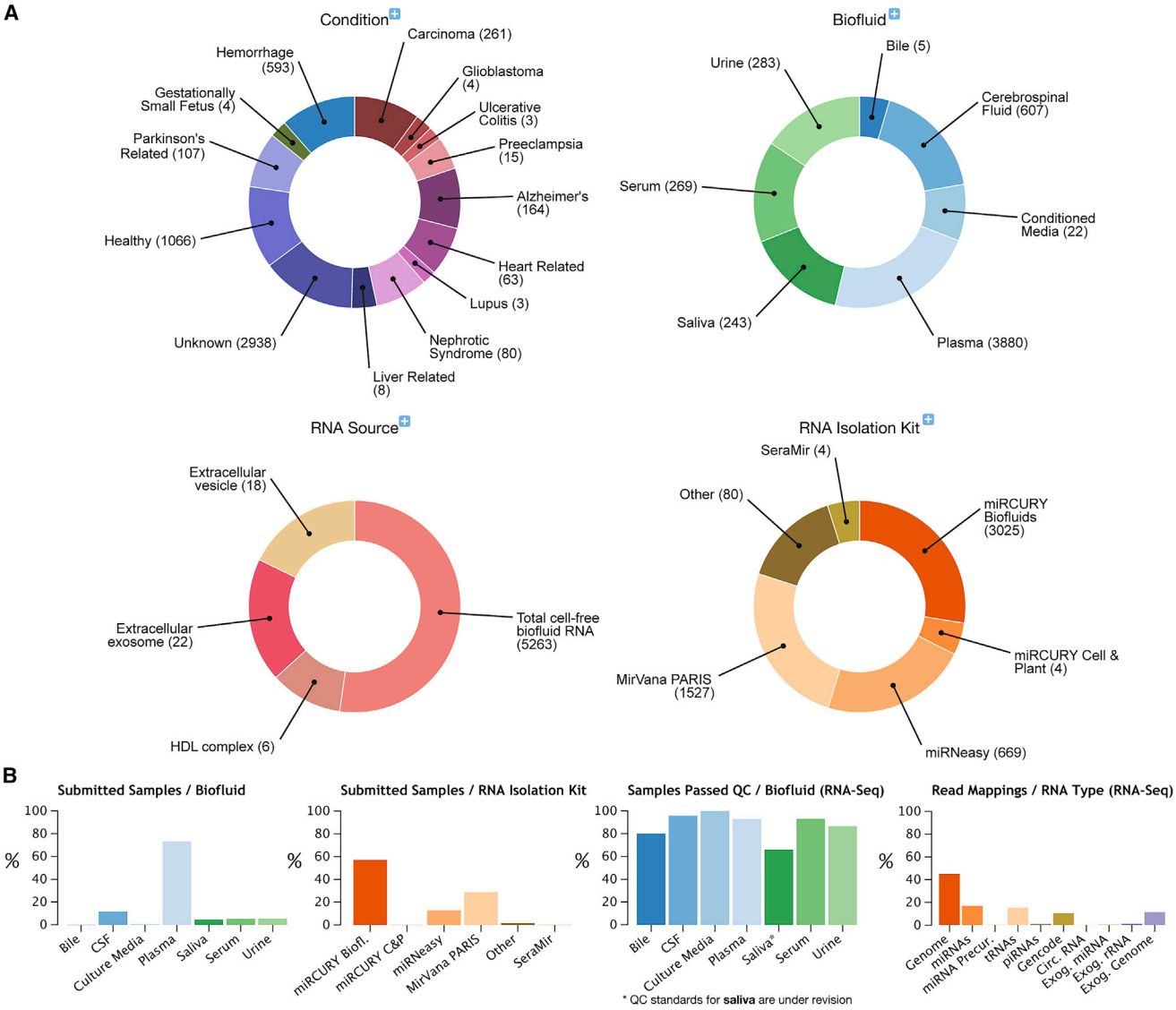


Figure 1. exRNA Atlas Resource

(A) Faceted charts for selecting exRNA profiles from the Atlas. The size of each slice (representing a profile count) has been log-transformed to aid usability. RNA source categories follow the protocols established by the ERCC.

(B) Bar charts describing the contents of the Atlas.

is available on the Atlas landing page (Figure 1B). The exRNA Atlas is dynamically populated using the extensive metadata collected and stored for exRNA profiling samples. Comprehensive metadata describing the sample, sample donor, protocols used to collect and prepare the sample for exRNA profiling experiments, and results of analytical methods are modeled using GenboreeKB. Many metadata fields employ ontologies such as the National Cancer Institute Thesaurus (NCIT) (Musen, 2013), SNOMED CT (Donnelly, 2006; Stearns et al., 2001), and Human Disease Ontology (DOID) (Kibbe et al., 2015) and are validated using the BioPortal API (Whetzel et al., 2011). An overview of the metadata standards developed by the ERCC with different metadata entities stored in the Atlas is available in Figure S1.

In addition to the graphical user interface (GUI) on the Atlas website, Atlas (meta)data are exposed via a dedicated JSON-LD Application Programming Interface (API) with full documentation available in standard OpenAPI format (https://exrna-atlas.docs.stoplight.io/). The Atlas resource is also indexed on Google Dataset Search and FAIRsharing.org (McQuilton et al., 2016). Additional details on navigating Atlas (meta)data can be found in STAR Methods.

Submitted RNA-Seq Profiles Are Uniformly Processed

Data flow into the exRNA Atlas is mediated by a data submission and processing pipeline (Figure 2A). The pipeline was designed to accept data from exRNA profiling experiments using RNA

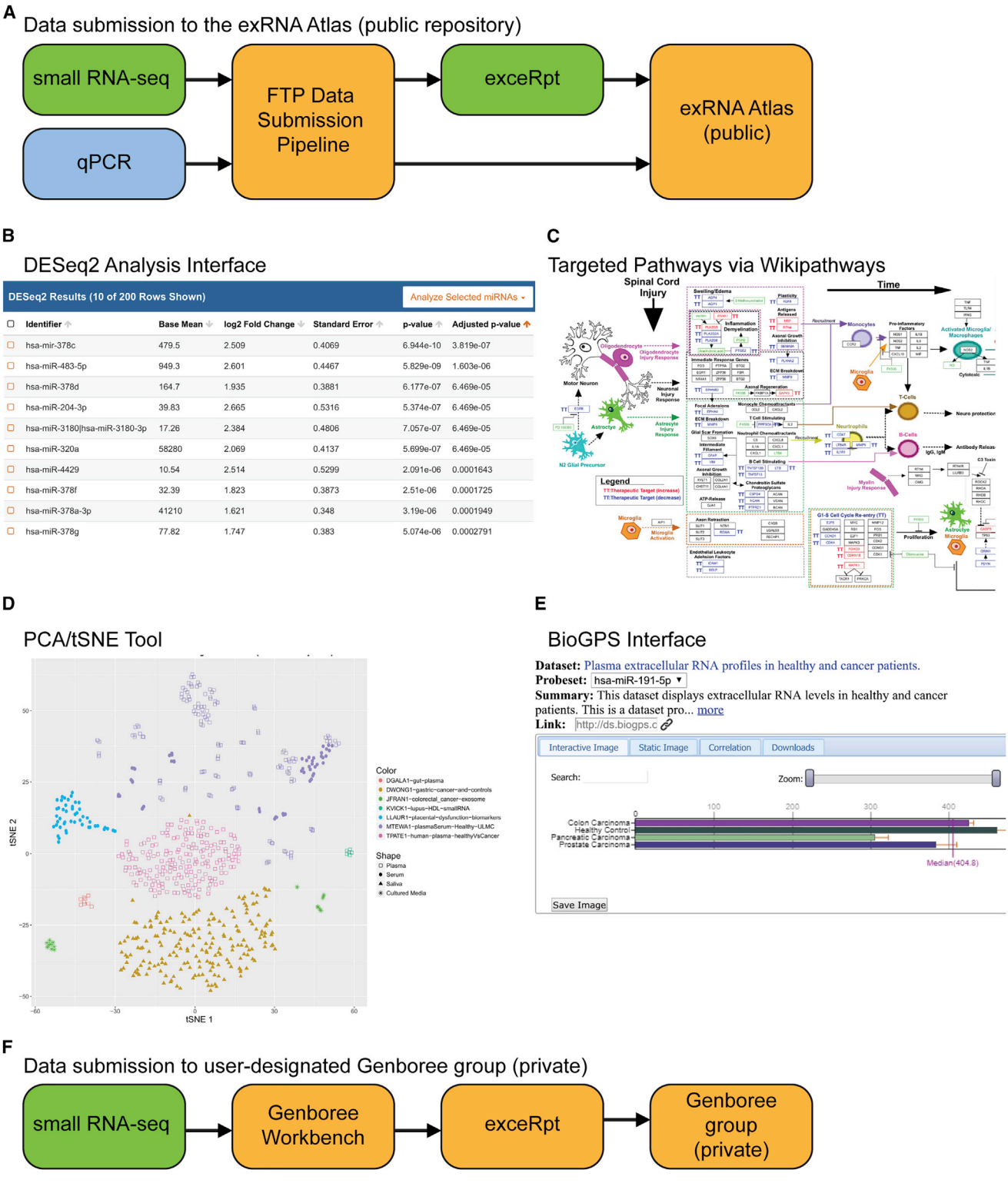


Figure 2. Overview of exRNA Atlas Data Submission Process and exRNA Atlas Tools

(A) Workflow for submitting exRNA profiling data to the Atlas. Submissions consist of three different file types: data files (optional for qPCR), metadata files, and a manifest file. All files are processed through an FTP-based data submission pipeline, with exRNA-seq data being uniformly processed through exceRpt. After validation, processing, and deployment, all data and metadata are made available through the Atlas website.

(legend continued on next page)

sequencing (RNA-seq) and qPCR. Detailed instructions guide users through the (meta)data preparation and submission process, thereby encouraging community contribution to the Atlas. Additional details on the data submission process are provided in STAR Methods. To minimize variability and facilitate integrative analyses across studies, all exRNA-seq data in the Atlas are uniformly processed using the extra-cellular RNA processing toolkit (exceRpt), an exRNA-seq processing pipeline created by members of the ERCC (Rozowsky et al., 2019). For Version 4P1 of the Atlas, a total of 23.43 billion reads from 2,270 RNA-seq sample profiles were processed through exceRpt. Reads from exRNA sequencing experiments are sequentially aligned to the host genome and transcriptome as well as to various exogenous genomes. Output from exceRpt includes abundance estimates for the various RNA libraries and detailed mapping information for each read mapped for each library, as well as a variety of metrics such as read-length distribution and summaries of reads mapped to each library. A set of quality measures agreed upon by members of the ERCC (Rozowsky et al., 2019) are generated for each profile, and the small number of profiles not passing QC thresholds are denoted. After all sample files are processed and deployed, data and metadata become available through the exRNA Atlas website.

exRNA Atlas Resource Includes Analysis and Visualization Tools

The exRNA Atlas' suite of available tools allows users to analyze existing exRNA-seq profiles in the Atlas, process their own exRNA-seq data, analyze their data in the context of the Atlas, and (at or before the time of publication) contribute their own data to the Atlas to empower future exRNA research. Pairwise differential expression analysis may be performed using DESeq2 (Love et al., 2014) (Figure 2B). Custom pathway queries for differentially expressed microRNAs (miRNAs) can be performed in the context of the Extracellular RNA section of WikiPathways (Slenter et al., 2018) via the Pathway Finder tool to provide a ranked list of targeted pathways (Figure 2C). Atlas studies may be visualized as precomputed principal component analysis (PCA) (Abdi and Williams, 2010) and t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction plots (van der Maaten and Hinton, 2008) (Figure 2D). Summary plots and tables may be generated for any set of exRNA-seq profiles (up to 900 at a time) from the Atlas. Users can visualize miRNA read expression for a given dataset via BioGPS (Wu et al., 2016), a gene annotation portal that contains interactive gene expression bar charts for Atlas exRNA-seq datasets (Figure 2E). Information on the expression of specific miRNA species across human biofluids is accessible via WikiData. Computational deconvolution results presented in this paper (described below) are available via the Atlas Public Analysis Results page. Finally, users can also analyze their own exRNA-seq data using the Genboree Workbench (Figure 2F), a web-based software platform which hosts several exRNA-seq-based bioinformatics tools (Amin et al., 2015). The Workbench allows users to upload and store their own exRNA-seq profiling data, process that data through the exceRpt pipeline, and share results privately with collaborators prior to sharing them publicly through the Atlas.

Computational Deconvolution Explains Most Variation across exRNA Atlas RNA-Seq Profiles

Atlas miRNA expression profiles cluster primarily by study, despite uniform processing through the exceRpt pipeline (Figure S2), suggesting large technical and possibly biological variability across studies. We reasoned that one potential source of variation may be variable representation of a multiplicity of exRNA carriers, each having a characteristic cargo profile. Supporting this contention, previous studies revealed highly distinct non-coding RNA (ncRNA) profiles for different extracellular RNA carriers in both human and mouse, including HDV and LDV extracellular vesicles isolated from the supernatant of the HMC-1 human mast cell line (Lässer et al., 2017) and lipoprotein profiles isolated from mice (Allen et al., 2018). We therefore hypothesized (1) that biologically meaningful invariant exRNA profile signatures exist across studies and possibly across biofluids, and (2) the invariant signatures may be overshadowed by the technical or biological variability in their relative proportions in individual samples and across studies and biofluids.

To test the hypothesis, we adapted a computational deconvolution method that we previously developed for highly heterogeneous human tumors (Onuchic et al., 2016). Based on the exRNA-seq profiles of whole biofluid or any of its fractions, the deconvolution algorithm estimates (1) the number of distinct constituent cargo profiles that are present—with possibly a small degree of sample-to-sample variation—in many samples, (2) the cargo profiles themselves, defined by the relative abundance of ncRNA species of any biotype (e.g., miRNA and large intergenic noncoding RNA [lincRNA]); and (3) the percent contribution of each cargo profile to the bulk profile of any given sample (Figure 3). One key requirement for deconvolution is the presence

(B) DESeq2 analysis interface. The integrated DESeq2 tool allows users to discover differentially expressed miRNAs in Atlas data via pairwise differential expression analysis. Users can launch their own analyses via the results grid or view precomputed analyses via the Public Analysis Results page.

(C) Pathway Enrichment via Pathway Finder on WikiPathways. Users can select miRNAs on the Atlas via their DESeq2 results or the Atlas census page and perform downstream pathway analyses via the Pathway Finder tool on WikiPathways. The Pathway Finder tool lists pathways that contain selected miRNAs and their targets.

(D) PCA/t-distributed stochastic neighbor embedding (t-SNE) tool interface. The integrated Dimensionality Reduction Plotting Tool allows users to visualize precomputed PCA/t-SNE analyses on Atlas datasets. All analyses are available via the Public Analysis Results page. See also Figure S2.

(E) BioGPS interface. Users can visualize individual miRNA read count expression via BioGPS for Atlas RNA-seq datasets. All Atlas BioGPS studies are available via the Datasets page.

(F) Workflow for submitting data to a Genboree group for personal analysis. Submissions consist of exRNA-seq data files (FASTQ) and are processed via exceRpt on the Genboree Workbench. After processing is completed, results can be shared privately with collaborators.

See also Figure S1.

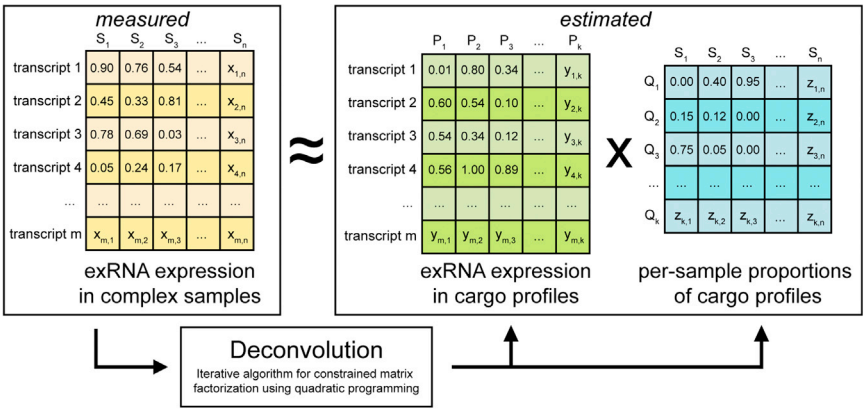


Figure 3. XDec Deconvolution Method

The exRNA expression profiles (transcript 1 to transcript m) of complex biofluid samples (S_1 to S_n) are used as the input. The deconvolution algorithm estimates the number (k) of constituent cargo profiles, the exRNA expression within the profiles (columns P_1 to P_k) and the proportions (rows Q_1 to Q_k) of the k cargo profiles in each sample (S_1 to S_n) through an iterative algorithm for constrained matrix factorization using quadratic programming. The algorithm involves two steps, the first involving transformed transcript abundance values over an informative set of ncRNAs and the second step involving non-transformed abundance values over all ncRNAs (STAR Methods). See also Figure S3 and STAR Methods.

of a sufficient number of ncRNA species that are informative for deconvolution. Two key criteria for a ncRNA species to be informative are (1) sufficient RNA-seq read coverage to ensure accurate quantitation, and (2) significant differences in the abundance between constituent exRNA cargo profiles. To generate a list of such ncRNAs, we compared RNA-seq profiles from previously characterized vesicular (HDV and LDV) (Lässer et al., 2017) and non-vesicular HDL (Atlas Dataset: EXR-KVICK1oIp40e-AN) carriers and identified 81 informative ncRNAs that show consistent quantitative differences between the carriers (Figure S3A; STAR Methods).

Deconvolution was applied to ncRNA profiles of bulk samples of 21 RNA-seq analysis datasets, totaling 2,138 samples (Table S1). Although exceRpt maps reads to both endogenous and exogenous ncRNA biotypes, we limited the exRNA Atlas biosample profiles to seven major endogenous RNA biotypes: miRNA, Piwi-interacting RNA (piRNA), tRNA, Y RNA, lincRNA, small nucleolar RNA (snoRNA), and small nuclear RNA (snRNA). Each of the 21 datasets represented a single disease state for a single biofluid and included at least 40 sample profiles. The datasets covered 5 biofluids (CSF, saliva, serum, plasma, and urine) and 9 disease states: healthy, gastric cancer, carcinoma (colon, prostate, pancreatic), Parkinson's disease, Alzheimer's disease, subarachnoid hemorrhage, intraventricular brain hemorrhage, myocardial infarction, and nephrotic syndrome (Table S1). For input to the deconvolution algorithm, reads per million mapped reads (RPM) values were transformed using quantile normalization. To eliminate overfitting of outliers and equalize measurements, the values were mapped to the [0,1] range using a negative exponential function (STAR Methods). The algorithm estimated the number of cargo profiles (k) independently for each of the 21 datasets using a stability criterion (STAR Methods). This resulted in k = 3 or 4 cargo profiles being estimated for each dataset, resulting in a total of 68 cargo profiles for the 21 datasets (Table S1), with an additional 7 profiles (for a total of 75) coming from physically isolated HDV, LDV, and HDL carriers.

We sought to quantify the level of variance explained by deconvolution. Specifically, we measured variance by multiple regression, approximating observed exRNA profiles by linear combinations of estimated profiles (k = 3 or 4 profiles estimated per dataset). In all datasets, between 50% and 90% of variance over the core set of informative ncRNAs was explained by deconvolution (Figure S3B, black point). The explained variance over the informative set was compared to the explained variance of 100 randomly selected ncRNA sets that matched in size and biotype (Figure S3B, boxplot). For 20 out of 21 datasets, the explained variance for the core set was significantly higher than for the randomly selected ncRNAs (Figure S3B; STAR Methods). However, for some studies, the variance of randomly selected ncRNAs was highly explainable, suggesting that many ncRNAs differ in abundance between cargo profiles. This led us to examine if read coverage is a significant determinant as to whether a ncRNA set is informative (Figure S3C). That indeed turned out to be the case, as indicated by the significant positive correlation ($p = 0.013$) between explained variance and the mean transformed read coverage of each study (Figure S3D).

Clustering of Study-Specific Deconvoluted Cargo Profiles Reveals Six Major Cargo Types

Next, we asked if the cargo profiles deconvoluted from one dataset show similarity to the profiles from other datasets. Such similarities may be expected to occur, for example, if a vesicular or non-vesicular carrier with a distinct exRNA cargo is present across different studies and biofluids. To check for the presence of such invariant profiles, we measured the pairwise correlations between the 75 profiles (68 cargo profiles, plus 2 HDV, 2 LDV, and 3 HDL profiles) and performed hierarchical clustering. The profiles clustered into six groups that we refer to as cargo types (CTs) denoted CT1, CT2, CT3A, CT3B, CT3C, and CT4 (Figure 4). Remarkably, all CTs were detected in at least two distinct biofluids. We note that, in some instances, a given CT was not detected in a particular biofluid; this may occur because of an actual absence of the CT in the specific biofluid or because of the incompleteness of the exRNA Atlas due to biases of specific RNA isolation kits utilized (discussed below). Finally, the number of samples available in some studies may have precluded detection of all cargo types represented. Therefore, further dataset integration using different RNA isolation kits and experimental designs may yield a more complete map of all cargo types across human biofluids.

Additionally, we combined the samples (n = 2,138) from the 21 datasets and applied deconvolution, anticipating this may yield a higher resolution map of cargo type heterogeneity. Indeed, while the deconvolution of individual datasets yielded only 3–4 cargo

profiles per dataset, the deconvolution of the combined set yielded 11 profiles (Figure 4), indicating deep CT heterogeneity. Moreover, each of the 11 profiles correlated with exactly one of the 6 CTs, suggesting that future accumulation of exRNA profiling data and their integrative analysis may uncover new subtypes of the six top-level CTs.

Carriers of Distinct Cargo Types Separate into Distinct Density Fractions

We assessed any differences in the density of carriers corresponding to the six CTs. Toward this goal we performed cushioned-density gradient ultracentrifugation (C-DGUC) of serum and plasma using OptiPrep density gradient from human donors (n = 5 male, n = 5 female). Four RNA samples, corresponding to three pools of OptiPrep fractions 1–3 (1.028–1.038 g/mL), 4–7 (1.046–1.079 g/mL), and 9–12 (1.106–1.259 g/mL), plus whole biofluid (serum and plasma), were extracted from each of the 10 donors for a total of 80 RNA samples. Two samples were then discarded due to low read count, yielding a total of 78 RNA-seq profiles (STAR Methods). The samples were prepared using the miRNeasy micro kit and NEBNext Multiplex small RNA Library preparation kit. This library preparation only allowed for capture of RNA fragments and not full-length versions of larger RNAs (e.g., lincRNA and Y RNA). Similar to the protocol identifying LDV and HDV, fraction 8 was excluded to avoid potential overlap between the high-density and low-density fractions of interest identified by sucrose gradient (Lässer et al., 2017).

Because the three OptiPrep fraction pools were selected based on the current knowledge about densities of known exRNA carriers, we hypothesized each fraction would be enriched for particular carriers and their cargo. To explore this hypothesis, we deconvoluted the 78 RNA-seq profiles, anticipating that the deconvoluted profiles would correspond to distinct cargo types. That indeed turned out to be the case based on the correlation pattern to the 75 profiles (Figure S4A), with a caveat that deconvoluted profile P1 (Figure S4A) turned out to correspond to CT4 with an additional component of CT1 and CT2. We note that this type of "limited resolution" is likely encountered in the deconvolution of some other exRNA Atlas datasets, as suggested by the absence of certain cargo types from some studies and biofluids and by the off-diagonal correlations in the heatmap (Figure 4).

Assuming there is a difference in the density of carriers of CT4, CT1, and CT2, the "resolution" problem may be overcome by increasing the number of profiles in the dataset (as evidenced by the high-resolution deconvolution of the complete exRNA Atlas discussed in the previous section). Another option, uniquely available in this dataset compared to Atlas datasets, was to identify a new set of informative ncRNAs (different from the original set of 81) by identifying ncRNA species that differ in abundance in the three density fractions. A new set of 80 ncRNAs selected based on that criterion and applied to the 78 serum and plasma samples yielded four deconvoluted profiles which correlated (correlation over the original set of 81 informative ncRNAs) with the four major cargo types (CT1–CT4) (Figures 4 and S5A). We observe significant correlation scores per profile across the 6 CTs (CT4 (fraction 1–3): $p < 2.2e{-}16$, CT1 (fraction 4–7): $p = 1e{-}11$, CT2 (fraction 9–12): $p = 2.7e{-}12$, CT3A–CT3C (whole biofluid): $p = 2e{-}15$) (Figure S5B; STAR Methods). Importantly, each profile was predicted to be in high proportion across one of the three density fractions and one corresponding to the whole biofluid (Figure S5C).

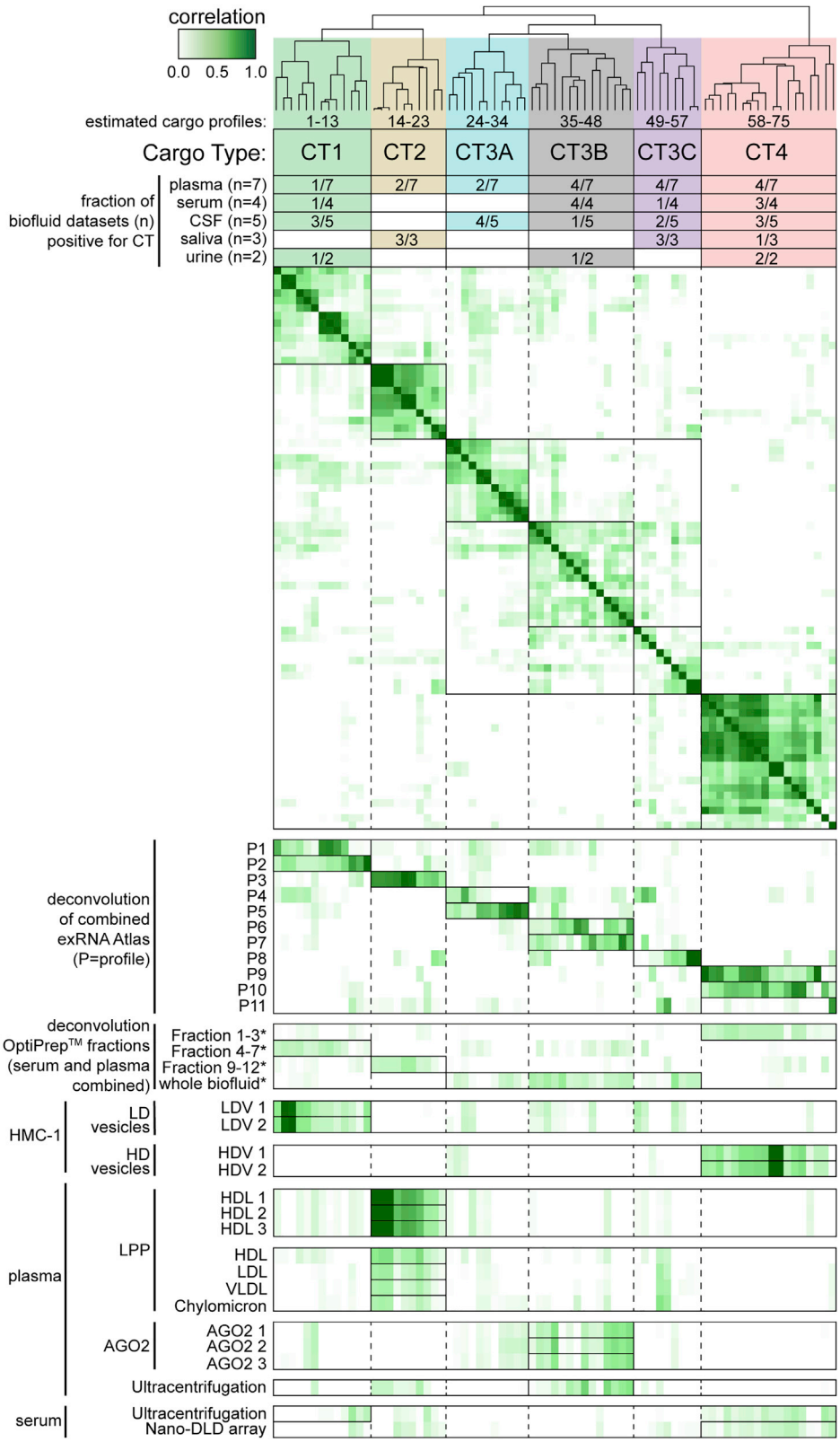
To better understand the informativeness of the new set of ncRNAs relative to the original set, we compared their biotype compositions. The majority of ncRNAs in the new set were lincRNAs, more than twice the number in the original set, with only a few miRNAs (Figure S4B). Moreover, none of the ncRNAs overlapped between the two sets. This divergent pattern may be explained by the different coverage of biotypes in the new dataset compared to Atlas datasets obtained using the same NEBNext small RNA library preparation kit (Figure S4C). The OptiPrep fractionation process may account for differences in RNA subtype abundances. Moreover, the read coverage of the new ncRNA set was negligible in the Atlas datasets (Figure S4D), highlighting the role of read coverage (that depends on sequencing library preparation) in determining informativeness of ncRNAs for deconvolution. This issue also indicates that deconvolution of the HDV, LDV, and HDL profiles would not resolve the constituent cargo types due to very low coverage of the new set of 80 informative ncRNAs (Figure S4D). Strikingly, despite using a completely new ncRNA set for deconvolution, and despite using density separation as the guiding principle, we obtained a one-to-one correspondence with the deconvoluted cargo types (Figure 4).

Cargo Types Associate with Known exRNA Carriers

We next sought to obtain information about the biological source of CTs using an additional array of methods for physical and immunological separation and characterization. Specifically, we correlated the 75 exRNA cargo profiles (columns in Figure 4) with RNA-seq profiles from (rows in the bottom half of Figure 4): (1) high density vesicles (HDVs) and low density vesicles (LDVs) from the HMC-1 human mast cell line (Lässer et al., 2017); (2) HDL particles isolated from plasma using immunoprecipitation (EXR-KVICK1oIp40e-AN); (3) purified lipoprotein particle (LPP) carriers (HDL, VLDL, LDL, chylomicron) isolated from plasma using sequential density ultracentrifugation (SD-UC) and fast-protein liquid chromatography (FPLC) (Li et al., 2018b); (4) AGO2-positive carriers obtained by immunoprecipitation from plasma (GSE124269); (5) pellets from plasma obtained by an ultracentrifugation (UC) protocol commonly used for isolating extracellular vesicles from cell supernatants (UC-plasma) (EXR-SADAS1EXER1-AN); (6) pellets from prostate cancer patients' serum obtained by a UC protocol (UC-serum) (Smith et al., 2018); and (7) extracellular vesicles isolated from prostate cancer patients' serum obtained utilizing nanoscale deterministic lateral displacement (nanoDLD) (Smith et al., 2018; Wunsch et al., 2016). Because we observed high heterogeneity within CTs (Figure 4), instead of using an average, we used the highest correlation to a member of a CT as an indication of CT membership. Below, we summarize these results and additional validation assays for each of the six CTs.

CT1 Associates with LD Vesicles

CT1 is detected in plasma, serum, CSF, and urine and correlates with OptiPrep fractions 4–7 (Figures 4 and S5A), corresponding to the density of extracellular vesicles. Serum and plasma fractions



(legend on next page)

4–7 are CD9 (25 kDa)- and flotillin (49 kDa)-positive by western blot (Figures S5E and S5F, respectively). Mass spectrometry of fractions 4–7 followed by pathway enrichment analysis (Figure S5H) revealed protein enrichments consistent with extracellular vesicles: in the biological process category, we detected exocytosis and secretion pathways, and in the cellular component category, we detected membrane-bounded vesicles, extracellular exosomes, and organelle lumen-parts. Consistent with an extracellular vesicle carrier, the CT1 cluster includes the LDV profiles (Figure 4). Additionally, extracellular vesicles isolated from serum by UC show the highest correlation with CT1 (Figures 4 and S6A). Deconvolution of these UC-serum samples resulted in the detection of a profile with high correlation with CT1 (Figure S6B) with the samples being primarily composed of that profile (Figure S6C). However, deconvolution identifies potential contamination of other RNA cargo types (CT3B and CT4) consistent with the detection of impurities by UC (Figure S6D) (Lobb et al., 2015).

CT2 Associates with Lipoproteins

CT2 is detected in plasma and saliva and correlates with OptiPrep fractions 9–12 (Figures 4 and S5A), consistent with the density of HDL. Western blotting showed that serum and plasma fractions 9–12 are positive for the HDL protein marker APOA1 (Figure S5G). Mass spectrometry of fractions 9–12 followed by pathway enrichment analysis revealed protein enrichments consistent with lipoproteins: in the biological process category, we detected enrichments in lipoprotein metabolic process and lipid metabolic process pathways, and in the cellular component category, we detected enrichment of spherical-HDL particle pathways (Figure S5H). Consistent with a lipoprotein carrier, CT2 includes the HDL profiles used to identify informative RNAs (Figure 4). Additionally, CT2 members show high correlation with profiles of an independent panel of lipoprotein carriers (HDL, LDL, VLDL, chylomicron) isolated using SD-UC and FPLC (Li et al., 2018b) (Figure 4). Previous reports have indicated differences in RNA cargo across lipoprotein subtypes in both human and mouse (Allen et al., 2018; Vickers et al., 2011); although deconvolution does not address this fine level of resolution, it does suggest that when considered in the broader context of non-lipoprotein cargo types, cargos of different lipoprotein carriers isolated from human plasma show high similarity (Figure S7A, LPP).

CT3A and CT3B Associate with AGO2-Positive Ribonucleoprotein, while CT3C Does Not

CT3A is detected in plasma and CSF, CT3B is detected in all biofluids except saliva, and CT3C is detected in all biofluids except urine (Figure 4). A significant fraction of these cargo types may escape pelleting due to high iodixanol levels and may therefore be overshadowed by other components within specific OptiPrep fractions. Members of CT3B show high correlation with the exRNA profiles of AGO2 immunoprecipitate while CT3A members show a moderate correlation indicating possible relation (Figure 4). CT3C does not show the same association, suggesting that CT3C may be contained by an AGO2-negative carrier or an AGO2-containing carrier where AGO2 is not accessible for immunoprecipitation. Additionally, plasma UC pellet correlates with CT3B (Figures 4 and S6A), indicating the presence of RNP particles in the pellet. In contrast, serum UC pellet does not correlate, consistent with experimental evidence that exRNAs trafficked by protein complexes in serum are destroyed or absorbed during the coagulation process (Max et al., 2018).

CT4 Associates with Vesicular Carriers of Variable Density

CT4 was shown to be the most distinct cargo type and is detected in all five biofluids; however, CT4 could not be definitively associated with a carrier of specific density (Figures 4 and 6SA). Paradoxically, CT4 correlates with OptiPrep fractions 1–3 (lowest density) from serum and plasma, while also correlating with much higher density fractions (HDVs) from the HMC-1 supernatant (Figure 4). One possible explanation for this paradox is that the previously recognized atypical content of exosomes derived from the HMC-1 cell line (Vukman et al., 2017) may account for their unusually high density compared to exosomal density in human body fluids. Consistent with vesicles being indeed present in the low-density fraction, mass spectrometry of fractions 1–3 followed by pathway enrichment analysis revealed protein enrichments consistent with vesicles: in the biological process category, we detected enrichment for vesicle-mediated transport, and in the cellular component category, we detected enrichment for cytoplasmic membrane-bounded vesicles and endocytic vesicles (Figure S5H). Additionally, we observed high correlation between vesicles (60 to 150 nm particles) purified using size exclusion nanoDLD technology from serum and CT4 (Figures 4 and S6A) (Smith et al., 2018; Wunsch et al., 2016). Deconvolution of these samples revealed a profile highly correlated with CT4 (Figure S6B) with the samples composed primarily of that profile (Figure S6C). Deconvolution also indicates that the nanoDLD vesicles are purified consistently and show negligible contamination from other cargo types based on per-sample CT proportions (Figure S6D).

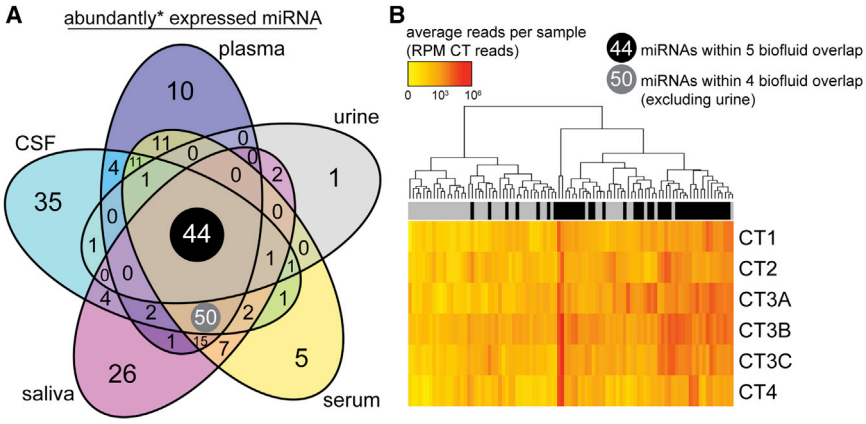
Cargo Types Show Distinct RNA Biotype Composition

We next quantified relative abundance of ncRNA biotypes for the six CTs by calculating the sum of all estimated reads per million (RPM) for each biotype across all members of each cargo type. Biotype proportions for CT1, CT2, and CT4 were similar to the proportions obtained from previous profiling studies of likely corresponding carriers (Figure S7A). CT3B-AGO2 has the largest proportion of miRNAs, consistent with AGO2 protein complexes being carriers of miRNA. The CT3B subtype also shows much

Figure 4. Deconvolution, Correlation, and Clustering of exRNA Atlas Datasets

Top self-heatmap represents the correlation scores of the 68 estimated cargo profiles, 2 HDV profiles, 2 LDV profiles, and 3 HDL profiles (75 total profiles) from the deconvolution of 21 individual analysis datasets in the exRNA Atlas. The top dendrogram represents the hierarchical clustering of the 75 profiles into 6 top-level clusters named Cargo Types: CT1, CT2, CT3A, CT3B, CT3C, and CT4. The table under Cargo Type names shows the biofluids where the cargo types are detected. The rows in the bottom of the figure show correlations of specific profiles (as indicated by the labels on the left) with the 75 profiles and Cargo Types. *Fractions are deconvoluted profiles corresponding to that given fraction. Additional details provided in Figure S5.

See also Figures S4–S7, Table S1, and STAR Methods.



*>10 mean RPM in at least 50% samples

Figure 5. Census Analysis of Abundant miRNAs

(A) Venn diagram representing the overlap between highly abundant miRNAs expressed at >10 mean RPMs in at least 50% samples for that biofluid within the Atlas. Black circle indicates the 44 miRNAs abundantly expressed within all 5 biofluids. Gray circle indicates the 50 miRNAs expressed within all biofluids except urine.

(B) Heatmap representing the RNA expression (log_{10}) level of the 94 highly abundant miRNAs across the predicted CT1–CT4. Color bar indicates if the miRNA was present in all 5 biofluid (black 44) or across 4 biofluids excluding urine (gray 50).

higher abundance of Y RNAs than other cargo types, consistent with previously observed abundance of Y RNAs in whole biofluids (Chakrabortty et al., 2015; Dhahbi et al., 2013; Yeri et al., 2017). These results should be considered in the context of large sample-to-sample variability in biotype proportions observed across Atlas studies (Figure S7B) and in the context of differences in library preparation protocols (Figure S5D).

Integration of exRNA Profiling Data Allows for a Census of Abundant miRNA

While the physiological role of miRNAs in human biofluids is still poorly understood, we reasoned that the miRNAs that are highly abundant in relevant biofluids may also be physiologically relevant. To identify such miRNAs for each biofluid, we developed a census of miRNAs present at >10 RPM (mean) in at least 50% of samples for that biofluid within the Atlas. A total of 44 miRNAs (Figure 5A, black), were expressed across all five biofluids and an additional 50 miRNAs (Figure 5A, gray) were expressed across four biofluids, but not urine. Complementing these results, a tool is available on the Atlas landing page to calculate a census of miRNAs and other ncRNAs based on user-selected thresholds using Atlas data. We also examined distribution of these 94 miRNAs across cargo types, which are represented at variable levels across the six CTs (Figure 5B).

Widely Used RNA Isolation Methods Show Cargo-Type Bias

The diversity of RNA isolation kits used in different studies is one potentially large source of variability. To assess potential kit biases for specific RNA content, an ERCC working group performed a multi-site study that evaluated 10 widely used RNA isolation methods used to extract RNA from a shared pool of human plasma and serum samples. The results of this analysis, as well as the original data, are available in the companion paper (Srinivasan et al., 2019).

As part of this collaborative effort, we assessed kit biases toward preferentially detecting specific CTs. For this purpose, we deconvoluted exRNA-seq profiles of RNA extracts obtained by applying different kits to identical plasma and serum samples (STAR Methods). The deconvolution algorithm estimated cargo profiles (k = 4) in the combined RNA-seq dataset using a stability criterion (STAR Methods). Correlation of the four profiles (across the 81 informative ncRNAs) to the six CTs classified two of them as belonging to CT1 and CT4, and the other two (referred as CT3B.1 and CT3B.2) as belonging to CT3B. Hierarchical clustering of the predicted per-sample proportions (Figure 6A, top heatmap) shows several sample clusters with similar composition based on carrier proportions. Clustering suggests several isolation methods are biased toward certain cargo types. However, per-sample proportion clustering does not reflect biofluid type. Additionally, certain sample clusters show similar gene expression profiles across observed groups of differentially expressed miRNAs (groups 1–4) (Srinivasan et al., 2019). These groups of miRNAs seem to be preferentially isolated by specific RNA isolation methods (Figure 6A, center heatmap). Particularly, samples with a relatively high proportion of CT1 have elevated levels of group 2 miRNAs. The estimated profile of CT1 (Figure 6A, left heatmap) indicates a similar expression pattern across the four miRNA groups. Furthermore, samples with a relatively high proportion of CT3B.1 have elevated levels of group 1 miRNAs; the CT3B.1 estimated profile shows similar miRNA expression.

Overall, the RNA isolation methods showed a diversity of preferences for specific CTs (Figure 6B). CT1 was captured in relative low abundance by all kits with ME and Millipore producing near zero amounts. CT4 was captured at highest relative abundance by all kits and was particularly enriched by ME, Millipore, and MiRCury kits. CT3B.1 was captured by all kits in small relative abundance and was highly enriched by miRNeasy. CT3B.2 was captured by all kits and was the overall second most abundant cargo type after CT4 (Figure 6B). We note that the deconvolution algorithm estimates only proportions of carrier RNA, not their absolute amounts, and a lower proportion of a specific CT does not imply lower absolute amounts of RNA.

Deconvolution of Plasma exRNA Profiles Detects Physiologically Relevant Pathway Signals

We next explored the potential of the deconvolution method to improve interpretation of case-control exRNA profiling studies. We reasoned that, by reducing variance, the method may help reveal biological signals that would otherwise be overshadowed by the variance. Moreover, by assigning any ncRNA differences between cases and controls to specific cargo

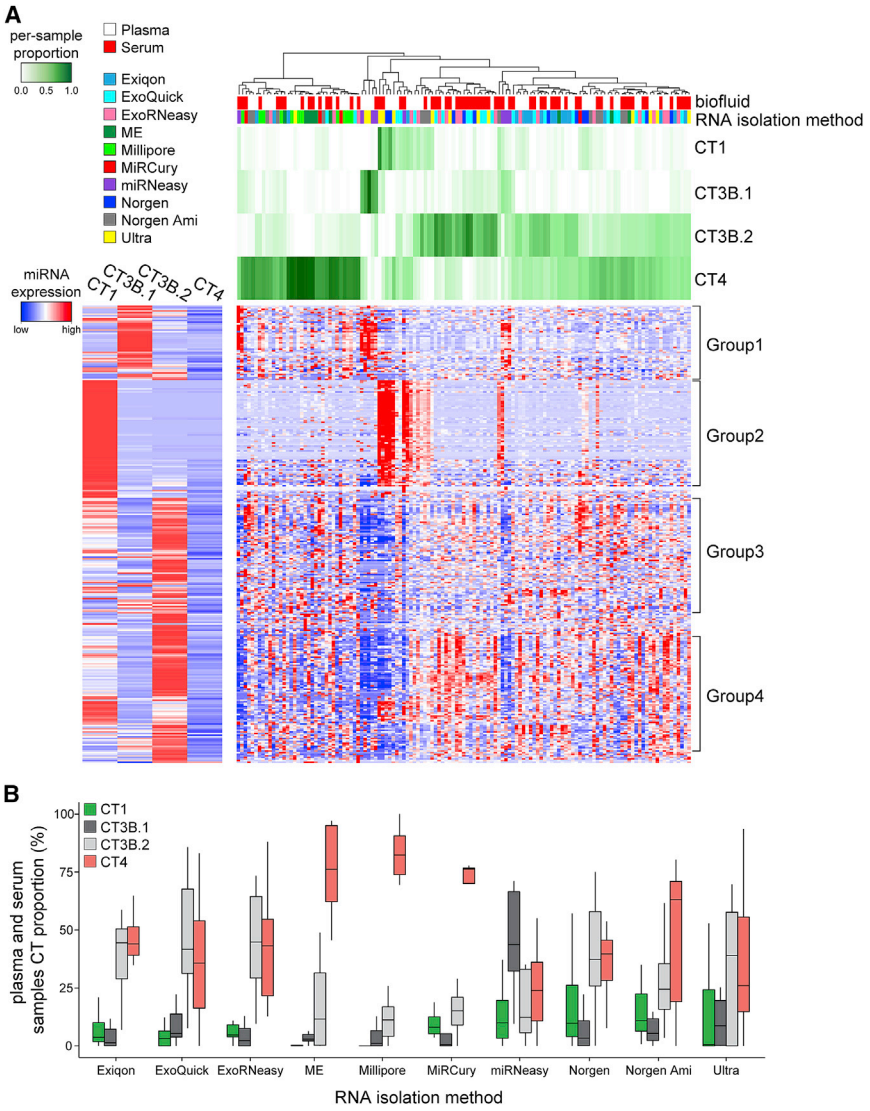


Figure 6. Deconvolution Estimates Cargo-Type Composition among RNA Isolation Methods

(A) Top heatmap shows hierarchical clustering of per-sample proportions of CTs predicted through computational deconvolution of plasma and serum biofluid samples. The biofluid and RNA isolation methods are color-coded above the heatmap. Center heatmap shows hierarchical clustering results of sample gene expression profiles (clustering of miRNAs was performed; samples are ordered based on dendrogram of per-sample proportions). groups 1–4, indicated to the right of the heatmap, are sets of miRNAs that are preferentially isolated by specific RNA isolation methods. Left heatmap shows expression profiles of the four CTs estimated through deconvolution across miRNA groups 1–4.

(B) Box-plot of per-sample proportions of four CTs estimated through computational deconvolution for all ten RNA isolation methods.

types, the method may provide deeper insights into the biology of any detected differences. Deconvolution of Atlas studies revealed that the proportions of CTs varied from sample to sample. In most case-control studies, however, there were no systematic differences in proportions between cases and controls (Figures S7C–S7F). One exception was the exercise challenge study (Shah et al., 2017), which compared pre- and post-exercise (Bruce treadmill test protocol) exRNA profiles of human plasma. For this dataset, deconvolution revealed pre- versus post-exercise differences in cargo type proportions and in ncRNA abundances within the cargo types. Because both types of differences could be demonstrated within this single study, we chose it as an example to comprehensively illustrate the power of deconvolution to detect relevant biological signals.

Deconvolution of the exercise study dataset revealed three cargo profiles (P1–P3) that were assigned to CT2, CT3B, and CT4 cargo types based on correlation across the 81 informative ncRNAs (Figure 7A). A significant ($p = 0.001$) increase in CT4 proportion was observed in the pre-exercise versus post-exercise group (Figure 7B). We note that because deconvolution estimates relative proportions of each component, an increase in one profile (CT4) is coupled to a reduction in the other components (CT2, CT3B); while change in CT4 abundance appears to be most prominent, simultaneous absolute changes in two or more components cannot be ruled out.

To detect miRNA differences pre- and post-exercise, we deconvoluted the pre- and post-exercise groups of profiles separately. A t test detected 53 differentially expressed miRNAs across the three CTs (Figure 7C), 17 (CT3B: 1, CT4: 16) of which were previously detected without deconvolution (Figure 7C, yellow circle). We identified pathways enriched for the differentially expressed miRNAs within each cargo type (Figure 7D). Strikingly, the four most significant pathways affected by the differentially expressed miRNAs within CT4 were those relating to striated muscle contraction and cell motility (Figure 7D, pink). CT2 showed miRNA changes consistent with an energy metabolism challenge (Figure 7D, yellow). Intriguingly, CT2 is associated with lipoprotein carriers, raising the question about the role for HDL and lipoproteins as conveyors of exRNA-mediated homeostatic responses to physical activity. In contrast to these findings, previously published analysis of the same dataset without deconvolution did not reveal any exRNA carriers or pathways that were specifically related to physical activity (Shah et al., 2017). Taken together, these results illustrate the potential of the deconvolution method to improve the detection and interpretation of potentially physiologically relevant exRNA perturbations.

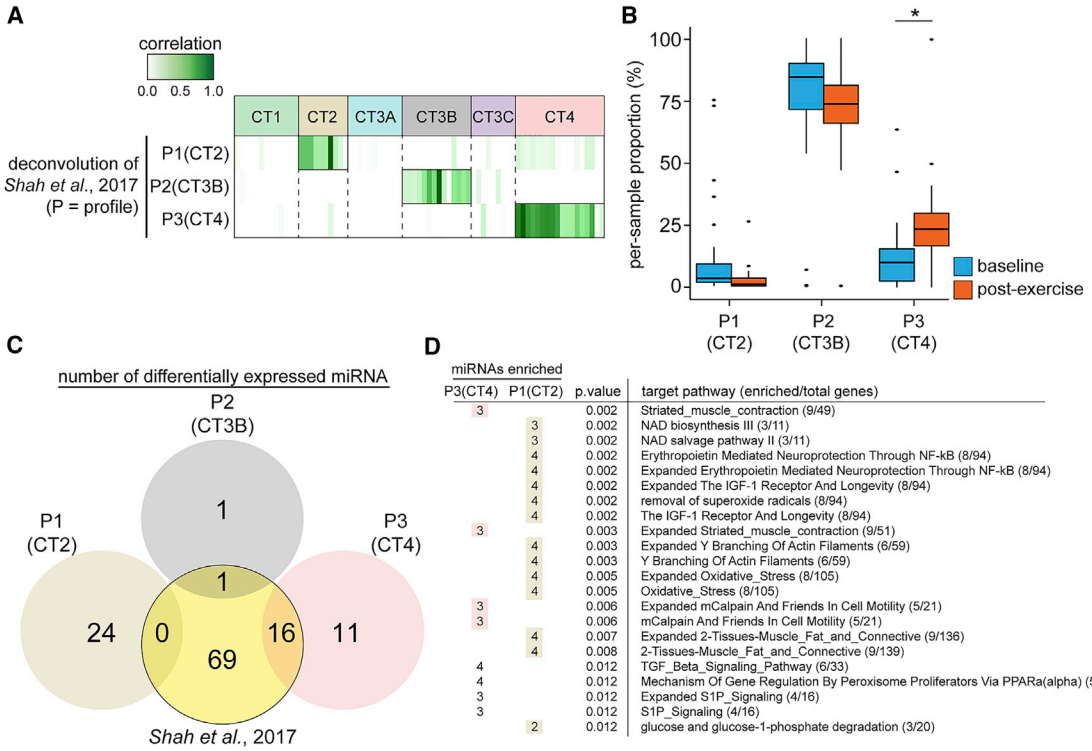


Figure 7. Deconvolution of Exercise Case Study

(A) Heatmap representing the correlation between the 3 cargo profiles modeled for Atlas Dataset: EXR-SADAS1EXER1-AN and the cargo profiles estimated from individual Atlas datasets that form the 6 CTs.

(B) Difference in abundance of each cargo profile between baseline samples and post-exercise samples (*p = 0.001). See also Figure S7.

(C) Number of differentially expressed miRNAs within each cargo profile. DESeq2 was used to identify differentially expressed miRNAs in the exceRpt-processed exercise dataset samples (yellow circle, Shah et al. [2017]). For methodological details see STAR Methods.

(D) mirnaPath was used to identify pathway enrichment for miRNAs differentially expressed for each cargo profile. Yellow highlighted boxes indicate pathways related to energy metabolism. Pink highlighted boxes indicate pathways related to muscle contraction and cell motility. For methodological details, see STAR Methods.

DISCUSSION

Deconvolutional meta-analysis of datasets within the Atlas reveals six major exRNA cargo types. Remarkably, the cargo types are detectable across diverse human biofluids. Five of the cargo types correspond to previously isolated and profiled vesicular and non-vesicular exRNA carriers. Taken together, these results constitute a milestone toward the construction of a map of extracellular RNA communication in humans. Our results also indicate that the heterogeneity of exRNA carriers and cargo types exceeds the capabilities of current experimental methods to reproducibly isolate and study defined carrier subpopulations and their cargo. While this problem clearly calls for the development of new carrier isolation methods, we have now demonstrated the power of computational deconvolution to complement and enhance such methods and tools.

While our findings suggest associations of cargo types with distinct carriers, other interpretations may not be definitively excluded. For example, RNA cargo may correspond to different mechanisms by which exRNA carriers are loaded and not fixed based on the carriers themselves. Moreover, the cargo types and loading mechanisms may vary by cell type. While future research will be required to address these possibilities, we anticipate that the cargo types inferred from the Atlas will provide a starting reference map that will inform and be refined by future studies.

Our study did not address directly the significant sequence-specific bias that appears to originate during library preparation and that may lead to up to four orders of magnitude differences in the depth of sampling of the same small RNA species by different library preparation protocols (Fuchs et al., 2015; Giraldez et al., 2018; Hafner et al., 2011; Hansen et al., 2010; Jayaprakash et al., 2011). These protocol-specific biases at least in part explain the fact that not all of the selected 81 ncRNAs are equally informative for deconvolution across all studies. The biases may also explain the large variation in relative amounts of ncRNA biotypes that we observe across different studies such as the striking contrast between the high abundance of lincRNAs in RNA-seq profiles obtained using most recent library preparation methods (Figure S5D) compared to their low abundance in the current version of the exRNA Atlas

(Figure S7A). Overall, however, our results suggest that the patterns required for deconvolution were not overshadowed by these biases, making cross-study comparison possible despite the differences between RNA isolation and library preparation protocols.

To enable wide application of deconvolution, the current Atlas pipeline combines the exceRpt pipeline with a deconvolution step. A number of options are available to potential users to use this combined pipeline: (1) direct Atlas submission, (2) web tool via the Genboree Workbench as a self-service for pre-publication analysis, and (3) private installation of the exceRpt pipeline and the deconvolution software. Up-to-date information about these resources is available at https://exrna-atlas.org/exat/tools/deconvolution.

In summary, our results provide the first outline toward a map of extracellular RNA communication in humans. We demonstrate the power of sharing exRNA data and tools through the exRNA Atlas resource to enhance interpretation of exRNA profiling data from individual studies as well as across studies. By catalyzing the virtual cycle of data sharing and knowledge creation, the Atlas resource is lowering the barriers toward the discovery of biological principles of extracellular RNA communication and their eventual translation into actionable biomarkers and exRNA therapies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - exRNA Atlas Sample Datasets
  - High-Density and Low-Density exRNA Profiles
  - HDL exRNA Profiles
  - Lipoprotein Particles exRNA Profiles
  - AGO2 exRNA Profiles
  - Plasma Ultracentrifugation exRNA Profiles
- HUMAN SUBJECTS
  - OptiPrep
  - Serum UC/nanoDLD
  - RNA Isolation Kits Study
- METHOD DETAILS
  - Data Submission to Atlas
  - exRNA Atlas data navigation and content accessibility
  - Physically Isolated OptiPrep Fractions
  - Ultracentrifugation and nanoDLD Samples
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - exceRpt Sequence Processing
  - Deconvolution
  - Pathway Analysis
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.02.018.

ACKNOWLEDGMENTS

This publication is part of the NIH Extracellular RNA Communication Consortium paper package and was supported by the NIH Common Fund's exRNA Communication Program. This work was supported by a grant from the Common Fund of the NIH (5U54 DA036134 to A.M., D.J.G., and M.B.G. and UH2/UH3 TR000923 to D.T.W.W., and UH2/UH3 TR000901 to S.D.). This work was also supported in part by the NCI, NIH (5RO1 CA163849 to A.M.); the NIH Initiative for Maximizing Student Development (2R25 GM056929 to O.D.M.); the NIH (UH3 TR000906 to S.S., C.D.L., and L.C.L.); the NIH (5U19 CA179512, HL133575, and P30 DK63720 to R.L.R.); the Intramural Research Program of the NCI, Center for Cancer Research, NIH (to J.A.W., L.M.J., and J.C.J.); NIH (K23-HL127099, R01-HL136685, and R01-AG059729 to R.S; NIH (UH3 TR000943, R35 CA209904, and CA217685 to A.K.S.); NIH (R01-HL 122547 to S.D.); the American Cancer Society Research Professor Award (to A.K.S.); and the Frank McGraw Memorial Chair in Cancer Research (to A.K.S.).

AUTHOR CONTRIBUTIONS

Conceptualization, O.D.M., W.T., J.R., S.L.S., R.L., L.C.L., M.G., and A.M.; Methodology, O.D.M., W.T., S.L.S., K-H.C., and A.M.; Software, O.D.M., W.T., J.R., S.L.S., R.L., N.S., A.R.J., R.R.K., T.G., J.W., J.A.D., K.H., A.R., C.W., R.P.A., A.R.P., A.I.S., and A.M.; Validation, S.S., A.C., C.D.L., R.L.R., J.A.W., L.M.J., J.C.J., M.E.A., C.C.-C., N.D., S.M.G., J.T.S., G.S., A.K.T., B.H.W., K.K.Y., K.M.D., and L.C.L.; Formal Analysis, O.D.M., W.T., and A.M.; Investigation, O.M., D.T.W.W., D.J.G., R.V.S., A.Y., K.V.K.-J., L.W., T.P., K.M.D., J.F., C.M., P.N., A.R.P., B.S., D.K.W., X.Z., L.B., R.G., A.K.S., R.L.R., S.D., K.-H.C., A.R.P., A.I.S., L.M.J., L.C.L., M.E.R., M.B.G., and A.M.; Resources, A.R.J., R.L.R., L.C.L., and A.M.; Data Curation, W.T., S.L.S., N.S., and S.B.-M.; Writing – Original Draft, O.D.M., W.T., S.L.S., and A.M.; Writing – Review & Editing, O.D.M., W.T., J.R., S.L.S., M.E.R., and A.M.; Visualization, O.D.M., W.T., S.L.S., A.R.J., J.A.D., and A.M.; Supervision, J.R., L.C.L., and A.M.; Project Administration, M.E.R., M.B.G., and A.M.; Funding Acquisition, O.D.M., J.A.W., L.M.J., J.C.J., R.L.R., M.E.R., M.B.G., and A.M.

DECLARATION OF INTERESTS

Within the last 12 months, R.V.S. has received funds from Amgen (scientific advisory board), Myokardia (consulting), and Best Doctors (consulting). A.K.T. has received funds from Proxamo (scientific advisory board) and Siemens (scientific advisory board). R.V.S. is a co-inventor on a patent for exRNAs signatures of cardiac remodeling. All other authors declare no competing interests.

Received: April 28, 2018
Revised: November 6, 2018
Accepted: February 11, 2019
Published: April 4, 2019

REFERENCES

Abdi, H., and Williams, L.J. (2010). Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. *2*, 433–459.

Ainsztein, A.M., Brooks, P.J., Dugan, V.G., Ganguly, A., Guo, M., Howcroft, T.K., Kelley, C.A., Kuo, L.S., Labosky, P.A., Lenzi, R., et al. (2015). The NIH Extracellular RNA Communication Consortium. J. Extracell. Vesicles *4*, 27493.

Amin, V., Harris, R.A., Onuchic, V., Jackson, A.R., Charnecki, T., Paithankar, S., Lakshmi Subramanian, S., Riehle, K., Coarfa, C., and Milosavljevic, A. (2015). Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. Nat. Commun. *6*, 6370.

Allen, R.M., Zhao, S., Ramirez Solano, M.A., Zhu, W., Michell, D.L., Wang, Y., Shyr, Y., Sethupathy, P., Linton, M.F., et al. (2018). Bioinformatic analysis of

endogenous and exogenous small RNAs on lipoproteins. J. Extracell Vesicles *7*, 1506198.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Burgos, K., Malenica, I., Metpally, R., Courtright, A., Rakela, B., Beach, T., Shill, H., Adler, C., Sabbagh, M., Villa, S., et al. (2014). Profiles of extracellular miRNA in cerebrospinal fluid and serum from patients with Alzheimer's and Parkinson's diseases correlate with disease status and features of pathology. PLoS ONE *9*, e94839.

Chakrabortty, S.K., Prakash, A., Nechooshtan, G., Hearn, S., and Gingeras, T.R. (2015). Extracellular vesicle-mediated transfer of processed and functional RNY5 RNA. RNA *21*, 1966–1979.

Cheung, K.H., Keerthikumar, S., Roncaglia, P., Subramanian, S.L., Roth, M.E., Samuel, M., Anand, S., Gangoda, L., Gould, S., Alexander, R., et al. (2016). Extending gene ontology in the context of extracellular RNA and vesicle communication. J. Biomed. Semantics *7*, 19.

Cogswell, J.P., Ward, J., Taylor, I.A., Waters, M., Shi, Y., Cannon, B., Kelnar, K., Kemppainen, J., Brown, D., Chen, C., et al. (2008). Identification of miRNA changes in Alzheimer's disease brain and CSF yields putative biomarkers and insights into disease pathways. J. Alzheimers Dis. *14*, 27–41.

Dhahbi, J.M., Spindler, S.R., Atamna, H., Boffelli, D., Mote, P., and Martin, D.I. (2013). 5′-YRNA fragments derived by processing of transcripts from specific YRNA genes and pseudogenes are abundant in human serum and plasma. Physiol. Genomics *45*, 990–998.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. Stud. Health Technol. Inform. *121*, 279–290.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Fuchs, R.T., Sun, Z., Zhuang, F., and Robb, G.B. (2015). Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. PLoS ONE *10*, e0126049.

Giraldez, M.D., Spengler, R.M., Etheridge, A., Godoy, P.M., Barczak, A.J., Srinivasan, S., De Hoff, P.L., Tanriverdi, K., Courtright, A., Lu, S., et al. (2018). Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. Nat. Biotechnol. *36*, 746–757.

Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J.T., Nusbaum, J.D., Morozov, P., Ludwig, J., et al. (2011). RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA *17*, 1697–1712.

Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. *38*, e131.

Jayaprakash, A.D., Jabado, O., Brown, B.D., and Sachidanandam, R. (2011). Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. *39*, e141.

Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., et al. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. *43* (Database issue, D1), D1071–D1078.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lässer, C., Shelke, G.V., Yeri, A., Kim, D.K., Crescitelli, R., Raimondo, S., Sjöstrand, M., Gho, Y.S., Van Keuren Jensen, K., and Lötvall, J. (2017). Two distinct extracellular RNA signatures released by a single cell type identified by microarray and next-generation sequencing. RNA Biol. *14*, 58–72.

Leinonen, R., Sugawara, H., and Shumway, M.; International Nucleotide Sequence Database Collaboration (2010). The sequence read archive. Nucleic Acids Res. *39* (*Suppl 1*), D19–D21.

Li, K., Wong, D.K., Hong, K.Y., and Raffai, R.L. (2018a). Cushioned–Density Gradient Ultracentrifugation (C-DGUC): A Refined and High Performance Method for the Isolation, Characterization, and Use of Exosomes. Methods Mol. Biol. *1740*, 69–83.

Li, K., Wong, D.K., Luk, F.S., Kim, R.Y., and Raffai, R.L. (2018b). Isolation of Plasma Lipoproteins as a Source of Extracellular RNA. Methods Mol. Biol. *1740*, 139–153.

Lobb, R.J., Becker, M., Wen, S.W., Wong, C.S., Wiegmans, A.P., Leimgruber, A., and Möller, A. (2015). Optimized exosome isolation protocol for cell culture supernatant and human plasma. J. Extracell. Vesicles *4*, 27031.

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

Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet. *39*, 1181–1186.

Max, K.E., Bertram, K., Akat, K.M., Bogardus, K.A., Li, J., Morozov, P., Ben-Dov, I.Z., Weiss, Z.R., Azizian, A., Sopeyin, A., et al. (2018). Human plasma and serum extracellular small RNA reference profiles and their clinical utility. Proc. Natl. Acad. Sci. USA *115*, E5334–E5343.

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., and Sansone, S.A. (2016). BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database (Oxford) *2016*, baw075.

Musen, M.A. (2013). National Cancer Institute Thesaurus. In Encyclopedia of Systems Biology, W. Dubitzky, O. Wolkenhauer, K.H. Cho, and H. Yokota, eds. (Springer).

Onuchic, V., Hartmaier, R.J., Boone, D.N., Samuels, M.L., Patel, R.Y., White, W.M., Garovic, V.D., Oesterreich, S., Roth, M.E., Lee, A.V., and Milosavljevic, A. (2016). Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. Cell Rep. *17*, 2075–2086.

Rozowsky, J., Kitchen, R., Park, J.J., Galeev, T.R., Diao, J., Warrell, J., Thistlethwaite, W., Subramanian, S.L., Milosavljevic, A., and Gerstein, M. (2019). exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. Cell Systems 8. https://doi.org/10.1016/j.cels.2019.03.004.other.

Shah, R., Yeri, A., Das, A., Courtright-Lim, A., Ziegler, O., Gervino, E., Ocel, J., Quintero-Pinzon, P., Wooster, L., Bailey, C.S., et al. (2017). Small RNA-seq during acute maximal exercise reveal RNAs involved in vascular inflammation and cardiometabolic health: brief report. Am. J. Physiol. Heart Circ. Physiol. *313*, H1162–H1167.

Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., et al. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. *46* (D1), D661–D667.

Smith, J.T., Wunsch, B.H., Dogra, N., Ahsen, M.E., Lee, K., Yadav, K.K., Weil, R., Pereira, M.A., Patel, J.V., Duch, E.A., et al. (2018). Integrated nanoscale deterministic lateral displacement arrays for separation of extracellular vesicles from clinically-relevant volumes of biological samples. Lab Chip *18*, 3913–3925.

Srinivasan, S., Yeri, A., Cheah, P.S., Chung, A., Danielson, K., DeHoff, P., Filant, J., Laurent, C.D., Laurent, L.D., et al. (2019). Small RNA Sequencing across Diverse Biofluids Identifies Optimal Methods for exRNA Isolation. Cell *177*, this issue, 446–462.

Stearns, M.Q., Price, C., Spackman, K.A., and Wang, A.Y. (2001). SNOMED clinical terms: overview of the development process and project status. Proc. AMIA Symp. *2001*, 662–666.

Subramanian, S.L., Kitchen, R.R., Alexander, R., Carter, B.S., Cheung, K.H., Laurent, L.C., Pico, A., Roberts, L.R., Roth, M.E., Rozowsky, J.S., et al. (2015). Integration of extracellular RNA profiling data using metadata, biomedical ontologies and Linked Data technologies. J. Extracell. Vesicles *4*, 27497.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43* (Database issue, D1), D447–D452.

Théry, C., Witwer, K.W., Aikawa, E., Alcaraz, M.J., Anderson, J.D., Andriantsitohaina, R., Antoniou, A., Arab, T., Archer, F., Atkin-Smith, G.K., et al. (2018). 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. J. Extracell. Vesicles *7*, 1535750.

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Vickers, K.C., Palmisano, B.T., Shoucri, B.M., Shamburek, R.D., and Remaley, A.T. (2011). MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. Nat. Cell Biol. *13*, 423–433.

Vukman, K.V., Försönits, A., Oszvald, Á., Tóth, E.A., and Buzás, E.I. (2017). Mast cell secretome: Soluble and vesicular components. Semin. Cell Dev. Biol. *67*, 65–73.

Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., and Musen, M.A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. *39* (Web Server issue, Suppl 2), W541-5.

Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

Wu, C., Jin, X., Tsueng, G., Afrasiabi, C., and Su, A.I. (2016). BioGPS: building your own mash-up of gene annotations and expression profiles. Nucleic Acids Res. *44* (D1), D313–D316.

Wunsch, B.H., Smith, J.T., Gifford, S.M., Wang, C., Brink, M., Bruce, R.L., Austin, R.H., Stolovitzky, G., and Astier, Y. (2016). Nanoscale lateral displacement arrays for the separation of exosomes and colloids down to 20 nm. Nat. Nanotechnol. *11*, 936–940.

Yeri, A., Courtright, A., Reiman, R., Carlson, E., Beecroft, T., Janss, A., Siniard, A., Richholt, R., Balak, C., Rozowsky, J., et al. (2017). Total extracellular small RNA profiles from plasma, saliva, and urine of healthy subjects. Sci. Rep. *7*, 44061.

Yuan, T., Huang, X., Woodcock, M., Du, M., Dittmar, R., Wang, Y., Tsai, S., Kohli, M., Boardman, L., Patel, T., and Wang, L. (2016). Plasma extracellular RNA profiles in healthy and cancer patients. Sci. Rep. *6*, 19413.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit monoclonal anti-CD9	Abcam	Cat# ab92726; RRID: AB_10561589
Rabbit monoclonal anti-Flotillin-1	Cell Signaling Technology	Cat# 18634; RRID: AB_2773040
Mouse monoclonal apoA-I Antibody (B-10)	Santa Cruz Biotechnology	Cat# sc-376818; RRID: AB_2797313
Mouse monoclonal IgGk BP-HRP	Santa Cruz Biotechnology	Cat# sc-516102; RRID: AB_2687626
Goat polyclonal IgG (H+L)	Thermo Fisher Scientific	Cat# A10547; RRID: AB_2534046
Critical Commercial Assays		
miRNeasy Kit	QIAGEN	217084
RNA Clean & Concentrator	Zymo Research	SKU R1013/4
OptiPrep	Sigma-Aldrich	SKU D1556
NanoDLD chips	Smith et al., 2018	N/A
Deposited Data		
Analyzed RNA-seq data for exRNA Atlas study: "Profiles of Extracellular RNA in Cerebrospinal Fluid and Plasma from Subarachnoid Hemorrhage Patients"	This paper	GEO: GSE121868
Raw and analyzed RNA-seq data for exRNA Atlas study: "Identifying novel small RNA biomarkers unique to patients with gastric cancer"	This paper	dbGaP: phs001767.v1.p1; GEO: GSE121870
Analyzed RNA-seq data for exRNA Atlas study: "ULMC Plasma and serum exRNA from healthy donors at University of Michigan"	This paper	GEO: GSE121869
Raw RNA-seq data for exRNA Atlas study: "Plasma extracellular RNA profiles in healthy and cancer patients"	Yuan et al., 2016	GEO: GSE71008
Raw and analyzed RNA-seq data for exRNA Atlas study: "Identifying urinary RNA as non-invasive biomarkers for progression of chronic kidney disease"	This paper	GEO: GSE121978
Analyzed RNA-seq data for exRNA Atlas study: "small RNA Sequencing of CSF Samples from Patients with IVH"	This paper	GEO: GSE121867
Analyzed RNA-seq data for exRNA Atlas study: "Small RNA-seq during acute maximal exercise reveal RNAs involved in vascular inflammation and cardiometabolic health"	This paper	GEO: GSE121874
Analyzed RNA-seq data for exRNA Atlas study: "Identifying novel small RNA biomarkers for electrical and mechanical remodeling post-MI (myocardial infarction)"	This paper	GEO: GSE121875
Raw and analyzed RNA-seq data for exRNA Atlas study: "High-Density Lipoproteins - small RNA Signatures in Systemic Erythematosus Lupus."	This paper	GEO: GSE121865
Raw RNA-seq data for exRNA Atlas study: "Total Extracellular Small RNA Profiles from Plasma, Saliva, and Urine of Healthy Subjects."	Yeri et al., 2017	dbGaP: phs001258.v1.p1
Raw RNA-seq data for exRNA Atlas study: "Profiles of Extracellular miRNA in Cerebrospinal Fluid and Serum from Patients with Alzheimer's and Parkinson's Diseases Correlate with Disease Status and Features of Pathology"	Burgos et al., 2014	dbGaP: phs000727.v1.p1
Raw RNA-seq data for isolated low-density (LD) and high-density (HD) exRNA profiles	Lässer et al., 2017	BioProject: PRJNA343960

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Raw and analyzed RNA-seq data for isolated lipoprotein particle (LPP) exRNA profiles	This paper	GEO: GSE124131
Raw and analyzed RNA-seq data for serum ultracentrifugation (UC) and nanoDLD exRNA profiles	This paper	GEO: GSE123736
Analyzed RNA-seq data for AGO2 exRNA profiles	This paper	GEO: GSE124269
Analyzed RNA-seq data for OptiPrep exRNA profiles	This paper	GEO: GSE123864
Analyzed RNA-seq data for RNA isolation exRNA profiles	This paper	GEO: GSE123865
Software and Algorithms		
Extracellular RNA Atlas	Subramanian et al., 2015	https://exrna-atlas.org
exceRpt	Rozowsky et al., 2019	https://github.gersteinlab.org/exceRpt/
DESeq2	Love et al., 2014	10.18129/B9.bioc.DESeq2
t-SNE	van der Maaten and Hinton, 2008	https://github.com/jdonaldson/rtsne/
PCA	Abdi and Williams, 2010	R package: stats
Expression Deconvolution (XDec)	This paper	https://github.com/BRL-BCM/XDec
Epigenomic Deconvolution (EDec)	Onuchic et al., 2016	https://github.com/BRL-BCM/EDec
mirnaPath	Cogswell et al., 2008	10.18129/B9.bioc.miRNApath
Pathway Finder	Slenter et al., 2018	https://www.wikipathways.org/index.php/WikiPathways
BioGPS	Wu et al., 2016	http://biogps.org/
Genboree Workbench	Amin et al., 2015	http://genboree.org/site/
STRING	Szklarczyk et al., 2015	https://string-db.org
Other		
FAIRsharing.org identifier for exRNA Atlas	McQuilton et al., 2016	FAIRsharing: biodbcore-001137

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Aleksandar Milosavljevic (amilosav@bcm.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

exRNA Atlas Sample Datasets

We utilized the following datasets from the exRNA Atlas (https://exrna-atlas.org): Accession ID: EXR-KJENS1WBaSro-AN (n = 523, GEO: GSE121868), EXR-KJENS1RID1-AN (n = 428, dbGaP: phs001258.v1.p1), EXR-KJENS1sPlvS2-AN (n = 345, dbGaP: phs000727.v1.p1), EXR-DWONG1qf3tcS-AN (n = 198, dbGaP: phs001767.v1.p1; GEO: GSE121870), EXR-MTEWA1cHYLo6-AN (n = 197, GEO: GSE121869), EXR-TPATE1OqELFf-AN (n = 192, GEO: GSE71008), EXR-MBITZ12SHVIr-AN (n = 80, GEO: GSE121978), EXR-KJENS12WGutU-AN (n = 70, GEO: GSE121867), EXR-SADAS1EXER1-AN (n = 62, GEO: GSE121874), EXR-SADAS1UJ0CzW-AN (n = 43, GEO: GSE121875), and EXR-KVICK1oIp40e-AN (n = 6, GEO: GSE121865). Note that GEO contains only processed data for EXR-KJENS1WBaSro-AN, EXR-DWONG1qf3tcS-AN, EXR-MTEWA1cHYLo6-AN, EXR-KJENS12WGutU-AN, EXR-SADAS1UJ0CzW-AN, and EXR-SADAS1UJ0CzW-AN - raw data for these datasets are either currently being processed for dbGaP or are undergoing IRB evaluation.

High-Density and Low-Density exRNA Profiles

We utilized the High-Density (SRX2191757, SRX2191758) and Low-Density samples (SRX2191759, SRX2191760) from Lässer et al. (2017). Raw data was downloaded from SRA (BioProjectID: PRJNA343960) and reprocessed with exceRpt.

HDL exRNA Profiles

We utilized the HDL (Healthy Control) small RNA-seq samples deposited into the exRNA Atlas by Kasey Vickers from Vanderbilt University School of Medicine (EXR-KVICK1oIp40e-AN, GEO: GSE121865). Samples were collected from plasma and RNA was isolated utilizing the miRNeasy (QIAGEN) kit.

Lipoprotein Particles exRNA Profiles

RNA-seq profiles of purified lipoprotein (LPP) carriers (HDL, VLDL, LDL, Chylomicron) were isolated from plasma using SD-UC and FPLC (Li et al., 2018b). Associated data (n = 4) can be found on GEO: GSE124131.

AGO2 exRNA Profiles

We utilized AGO2 small RNA-Seq profiles isolated using anti-AGO2 immunoprecipitation from plasma from healthy human controls. Associated processed data (n = 3) can be found on GEO: GSE124269. Related raw data is currently undergoing IRB evaluation and will be released on dbGaP once evaluation is complete.

Plasma Ultracentrifugation exRNA Profiles

We utilized the extracellular vesicles (Healthy Control) small RNA-seq samples deposited into the exRNA Atlas by Saumya Das (EXR-SADAS1EXER1-AN, GEO: GSE121874). Samples were collected from plasma and RNA was isolated utilizing the MiRVana Paris (Ambion) kit.

HUMAN SUBJECTS

OptiPrep

For the OptiPrep samples, 250 mL whole blood was collected from five adult female and five adult male consenting donors supervised by Dr. Louise C. Laurent at the University of California, San Diego, Department of Obstetrics, Gynecology, and Reproductive Sciences and Sanford Consortium for Regenerative Medicine. These samples were utilized to generate the serum and plasma OptiPrep fractions and whole biofluid. Associated processed data (n = 78) can be found on GEO: GSE123864. Related raw data is currently undergoing IRB evaluation and will be released on dbGaP once evaluation is complete.

Serum UC/nanoDLD

Whole blood samples (2 to 5 ml) were collected by the team of Dr. Ashutosh Tewari at the Icahn School of Medicine at Mount Sinai, New York, Department of Urology by venipuncture from 9 consenting adult male Prostate Cancer patients under Institute Review Board approved protocols (GCO # 06-0996, 14-0318, and surgical consent) in purple capped tubes. After blood collection, serum was isolated using BD Vacutainer blood collection tubes, serum separation tubes (Fisher Scientific, Cat # 368016) and kept at −80°C until further steps were taken for exosome isolation. Serum was rapidly thawed prior to EV isolation with both nanoDLD and UC. Associated data (n = 14) can be found on GEO: GSE123736.

RNA Isolation Kits Study

Associated processed data (n = 182) can be found on GEO: GSE123865. Related raw data is currently undergoing IRB evaluation and will be released on dbGaP once evaluation is complete.

METHOD DETAILS

Data Submission to Atlas

Each submitting lab is assigned a dedicated area on the FTP server where submissions can be uploaded. Users upload three types of files: a data archive, a metadata archive, and a manifest file. For RNA-seq data, the data archive contains all sample sequencing files as well as an optional oligo spike-in file. For qPCR data, the data archive contains raw target value files. The metadata archive contains a selection of different tab-delimited files, each describing metadata associated with some entity. Different metadata entities include: Submissions, Studies, Runs, Biosamples, Experiments, Donors, Analyses, and qPCR Targets. All submitted metadata are validated against the relevant models stored in GenboreeKB. Most entities are required for submission, but Analyses entities are automatically generated by the data submission pipeline. The qPCR Targets entity is only required if the user is submitting qPCR data. The manifest file provides important supplementary information for the submission—for instance, for RNA-seq data submissions, it connects each sample data file with its respective biosample metadata.

Upon detecting a complete set of new submission files in one of the assigned areas, a monitor validates the submission content on a step-by-step basis to ensure correctness and integrity of the content and either (a) notifies the submitter via email about any detected problems, together with detailed instructions on how to fix them, or (b) arranges to run the processing pipeline on a batch execution cluster. The pipeline is a multi-phase, multi-job workflow with several parallel execution phases. The monitor creates a job plan for the pipeline workflow, pre-determining any job-dependencies. The job plans employ the more typical moderate resource level, so as not to waste computational resources and to increase the degree of parallelism. However, jobs within parallel phases do occasionally, if rarely, fail due to insufficient resources, such as RAM. The pipeline will automatically re-run such jobs with more resources; furthermore, it will update the job-dependencies in the plan initially created by the monitor. The parallel execution phase includes running the exceRpt pipeline on each sample or mapping to all exogenous genomes. Upon successful completion of the pipeline, the raw data and processed results are moved to dedicated storage areas, and accompanying metadata for the processed samples are stored in GenboreeKB. Depending on the restrictions defined by the ERCC Data Sharing and Access Policy

(https://exrna.org/resources/data/data-access-policy-summary/), relevant files are also made available in the public Genboree FTP server for download through the Atlas.

exRNA Atlas data navigation and content accessibility

The exRNA Atlas provides users with the ability to effectively find and navigate sample profiles using a rich set of metadata. The Atlas site supports four different search methods. First, the faceted charts on the landing page allow users to select profiles based on a combination of health conditions, biofluid, exRNA source, and/or RNA isolation experimental methods (Figure 1A). Second, two different biosample partition grids allow users to select cross-sections of Atlas data based on a biofluid versus health condition as well as a biofluid versus assay type. Third, users can select specific sample profiles via a tree selector that branches based on anatomical location, biofluid, and health condition. Finally, users can visit the Datasets page to view sample profiles associated with a particular study of interest. Each dataset is represented as a ''card'' which links to a metadata-centric view of the associated sample profiles. Users can access various links within the dataset cards to view associated publications (PubMed) and dataset pages on external public data repositories such as GEO (Edgar et al., 2002), SRA (Leinonen et al., 2010), and dbGaP (Mailman et al., 2007), and users can also download exceRpt summary files for exRNA-seq datasets. Summary files include a series of plots, including read count distributions, heatmaps for fraction of aligned reads for each alignment step of exceRpt, QC results, biotype distributions and read counts, miRNA abundance distributions, and exogenous genomic taxonomy hierarchical clustering plots. Each exRNA-seq profile in the Atlas also has an associated ''core results'' archive. This archive contains read count information on an RNA-species level for all of the RNA libraries, quality pass/failure status, and NCBI taxonomy trees generated from exogenous ribosomal RNA and genomic reads. If data associated with a given sample have no data use restrictions, original sequence data (FASTQ) as well as read alignment files (BAM) from the various alignment steps detailed above are also provided for download. Each qPCR profile in the Atlas contains a list of target ncRNAs and their Ct values. Users can access the experimental protocol metadata for any given RNA-seq or qPCR profile to learn more about the specific techniques used to generate that profile.

Physically Isolated OptiPrep Fractions

Serum and Plasma Samples

250 mL whole blood was collected from five adult female and five adult male donors using 19G needles using 60 mL syringes containing either no additive (for serum) or 440 uL 0.5 M K2EDTA pH 8. The blood was then transferred into 50 mL polypropylene tubes and allowed to sit at room temperature 10-60 minutes. The tubes were then spun at 2,000 xg for 20 minutes at room temperature. The clear supernatant was transferred to a fresh tube and centrifuged again at 2000 xg for 10 min at room temperature. The serum or plasma was then aliquoted into 1.5ml tubes and stored at −80°C.

Fractionation of Plasma & Serum Using Cushioned Density Gradient Ultracentrifugation (C-DGUC)

A volume of 0.8 mL of each serum and plasma sample was individually mixed with 39 mL of PBS, placed into an ultracentrifuge tube with a nominal capacity of 39 mL, and underlaid with 2 mL 60% iodixanol. The tubes were spun at 100,000 xg for 2 hours at 4°C. The bottom 3 mL (2 mL iodixanol cushion + 1 mL supernatant was removed, mixed, and underlaid under a step gradient of iodixanol (5%–10%-20% iodixanol in 0.25 M sucrose, 1 mM EDTA, and 10 mM Tris-HCl, pH 7.4). This was spun at 100,000 xg for 18 hours at 4°C. Twelve 1 mL fractions were then collected, starting from the top of the gradient as recently described (Li et al., 2018a). This included OptiPrep fractions 1-3 (1.028 - 1.038 g/mL), 4-7 (1.046 - 1.079 g/mL), and 9-12 (1.106 - 1.259 g/mL). The refractive index was measured using the RBD-6000 Series Refractometer (LAXCO) and the conversion to density was determined based on standard curve with 10/20/40/60% Iodixanol.

RNA Preparation

RNA from unfractionated serum and plasma samples (500 uL each) was isolated using the miRNeasy micro kit (QIAGEN) and concentrated using a Zymo RNA clean and concentrator-5 kit with a final elution volume of 7 uL. From each OptiPrep gradient, we combined fractions 1-3 (numbered from the top of the gradient) to form the light fraction, fractions 4-7 for the low-density fraction, and 9-12 for the high-density fraction. RNA was isolated from 500 uL of each of these combined fractions using the miRNeasy micro kit (QIAGEN) and concentrated using a Zymo RNA clean and concentrator-5 kit with a final elution volume of 7 uL.

Sequencing

4 uL of each RNA sample was dried in a SpeedVacTM. The dried RNA was then resuspended in 1.2 uL of water and used to generate a small RNA-seq library using the NEBNext Multiplex small RNA Library Prep kit. The library reactions were performed at $^1/_5$ scale using a Mosquito HTS liquid handler. The 80 libraries were combined into 2 pools, which were size-selected using a Pippen Prep with a cutoff of 117-180 bp. Each size-selected pool was run on one lane of a HiSeq 4000.

Western Blot Analysis

37.5uL of pool and 1uL of serum were resolved on a 10% gel and transferred onto PVDF. The membrane was blocked with 5% nonfat milk for 1 hour at room temperature. Primary antibodies were diluted in 1% milk and blots were probed overnight at 4°C. The blot was washed four times for five minutes with 0.1% PBST and incubated with either anti-mouse IgG HRP (Santa Cruz) or anti-rabbit IgG HRP (Thermo Fisher) at a dilution of 1:1000 for 1 hour at room temperature. The membrane was washed four times for five minutes with 0.1% PBST and rinsed with PBS. Blots were incubated with Amersham ECL Prime (GE Life Sciences) and imaged using the ImageQuant LAS 4000.

Mass Spectrometry Analysis

Patient serum samples were first immunodepleted of abundant serum proteins using the Human-14 Multiple Affinity Removal System column, according to manufacturer's instructions (Agilent). Immunodepleted serum samples or fractionated samples were in-solution digested with trypsin using the filter aided sample preparation (FASP) method (Wiśniewski et al., 2009). The resultant peptides were desalted using a Sep-Pak cartridge (Waters) and dried. For mass spectrometry analysis, all peptides were trapped on a trapping column and separated on a 75 μm x 15 cm, 2 μm Acclaim PepMap reverse phase column (Thermo Scientific) using an UltiMate 3000 RSLCnano HPLC (Thermo Scientific). Peptides were separated at a flow rate of 300 nL/min followed by online analysis by tandem mass spectrometry using a Thermo Orbitrap Fusion mass spectrometer. Peptides were eluted into the mass spectrometer using a linear gradient from 96% mobile phase A (0.1% formic acid in water) to 55% mobile phase B (0.1% formic acid in acetonitrile) over 30 minutes. Parent full-scan mass spectra were collected in the Orbitrap mass analyzer set to acquire data at 120,000 FWHM resolution; ions were then isolated in the quadrupole mass filter, fragmented within the HCD cell (HCD normalized energy 32%, stepped ± 3%), and the product ions analyzed in the ion trap. Proteome Discoverer 2.2 (Thermo) was used to search the data against human proteins from the UniProt database using SequestHT. The search was limited to tryptic peptides, with maximally two missed cleavages allowed. Cysteine carbamidomethylation was set as a fixed modification, and methionine oxidation set as a variable modification. The precursor mass tolerance was 10 ppm, and the fragment mass tolerance was 0.6 Da. The Percolator node was used to score and rank peptide matches using a 1% false discovery rate.

Ultracentrifugation and nanoDLD Samples

Serum Samples

Whole blood samples (2 to 5 ml) were collected by the team of Dr. Ashutosh Tewari at the Icahn School of Medicine at Mount Sinai, New York, Department of Urology by venipuncture from 9 consenting adult male Prostate Cancer patients under Institute Review Board approved protocols (GCO # 06-0996, 14-0318, and surgical consent) in purple capped tubes. After blood collection, serum was isolated using BD Vacutainer blood collection tubes, serum separation tubes (Fisher Scientific, Cat # 368016) and kept at −80°C until further steps were taken for EV isolation. Serum was rapidly thawed prior to EV isolation with both nanoDLD and UC.

nanoDLD Isolation of EVs from Serum

NanoDLD chips were fabricated in 200 mm silicon wafers, diced into individual chips, wetted in DI water (Millipore) and primed with 5% bovine serum albumin (BSA) to reduce non-specific adsorption and fouling as described in Smith et al. (2018). Individual chips were placed in custom-built acrylic flow cells prior to running serum samples. Post filtered (Whatman 0.2 um filters) serum samples (500 ul) were processed using nanoDLD chips with a gap of G = 150 nm at a flow rate of 4 μl/min for 60 min. Sample fluid enriched in EV's of size between 60 nm and 150 nm (as determined by EM and NTA) was removed from the nanoDLD bump outlet. Samples were stored at 4°C for RNA-seq processing.

UC Isolation of EVs from Serum

Serum samples were centrifuged in 5 mL tubes for 30 min at 2,000 g, 4°C. Supernatant was transferred into 50 mL tubes and tubes were filled with PBS up to $^{3}/_{4}$ levels of total volume, for 45 min of centrifugation at 12,000 g, 4°C. The supernatant was carefully transferred to UC tubes (Beckman coulter, thick wall polypropylene tube, Cat # 355642), and centrifuged for 2 hours at 110,000 g, 4°C. Pellets were resuspended in 1 mL PBS and UC tubes were filled up to $^{3}/_{4}$ of total capacity. After another round of centrifugation for 2 hours at 110,000, 4°C, pellets were resuspended in 1 mL PBS and stored at −80°C.

RNA Preparation

Total RNA was extracted from EV-enriched nanoDLD and UC processed serum using the Total Exosome RNA and Protein Isolation Kit (Invitrogen 4478545). RNA quality was assessed by bioanalyzer (Agilent 2100 Bioanalyzer, RNA 6000 Pico Kit, Agilent Technologies) and stored at −20°C. cDNA Libraries were prepared for small RNAs using the SMARTer smRNA-seq Kit for Illumina (Takara Bio 635030). Final library quality was verified with Qbit and bioanalyzer.

Sequencing

Next-generation RNA sequencing was performed using a HiSeq 4000 (Illumina), 100 base pair, single end reads at the New York Genome Center.

QUANTIFICATION AND STATISTICAL ANALYSIS

exceRpt Sequence Processing

The extra-cellular RNA processing toolkit (exceRpt) was developed for the processing and analysis of RNA-seq data generated to profile small exRNAs (Rozowsky et al., 2019). The pipeline is highly modular, allowing the user to define the libraries containing small RNA sequences that are used during RNA-seq read-mapping, and includes an option to provide a library of spike-in sequences to permit absolute quantitation of small-RNA molecules. The pipeline first performs a series of pre-processing and filtering steps designed to remove contaminants and prepare the samples for processing by automatic detection and removal of 3′ adaptor sequences, sequences that map to a pre-specified oligo spike-in library, and 45S, 5S, and mitochondrial rRNAs. Next, the remaining reads are aligned in parallel to the host genome and transcriptome, including mapping to miRNAs, tRNAs, piRNAs, GENCODE annotations, and circular RNAs. Finally, any remaining reads are aligned to exogenous miRNAs and rRNAs, and then to full genomes of bacteria, plants, fungi, protists, viruses, and certain vertebrates expected to be present in a human and/or mouse diet (Rozowsky

et al., 2019). All alignments in exceRpt are performed using the Spliced Transcripts Alignment to a Reference (STAR) software (Dobin et al., 2013), with the exception of alignment to oligo spike-in libraries, which is performed using Bowtie 2 (Langmead and Salzberg, 2012). After the pipeline finishes processing all submitted samples, a separate post-processing tool (Generate Summary Report) is run on all successful pipeline outputs. This tool generates useful summary plots and tables that can be used to compare and contrast different samples.

Deconvolution

The deconvolution of Onuchic et al. (2016) was first adapted to exRNA-seq data and then applied to the exRNA Atlas datasets (n = 21) listed in Table S1. Where appropriate, the Figure Legends detail the statistical test and paraments used to analyze the data corresponding to that Figure.

Gene Expression Transformation

For the first stage of the deconvolution algorithm, reads per million mapped reads (RPM) values are transformed using quantile normalization across each ncRNA independently. Additionally, expression values are fit to a range of [0,1] using negative exponential modeling: $1 - e^{-ax}, a = 1/max(RNA\ expression)$

This transformation ensured that the values will be in the 0-1 range, similarly to beta methylation values in the original algorithm and thus usable in the Stage 1 of deconvolution.

Identifying Informative ncRNAs for HD Vesicles, LD Vesicles, and HDL Protein Complexes

We began with our set of references that include samples from each group: HD vesicles, LD vesicles (Lässer et al., 2017) and HDL protein complexes (Atlas ID: EXR-KVICK1oIp40e-AN). To select informative ncRNAs, we performed t tests comparing the transformed expression levels over each RNA between each group of references against the rest of the reference profiles. We selected those RNAs that showed significant differences (p value = 0.000015) in the comparison of each group against the rest of the reference samples. Due to the greater similarity between HD vesicles and LD vesicles, we performed a specific t test comparing only the samples in those two groups and included in our final set of probes those that had a significant difference (p value < 0.003). Because of overlap between the RNA sets in each comparison, the final set contained 81 informative RNAs.

Identifying Informative ncRNAs for OptiPrep Fractions

Informative ncRNAs were identified for a set of independent RNA-seq profiles from OptiPrep fractionated serum/plasma samples. Based on the OptiPrep fractionation, we pooled fractions 1-3 (1.028 - 1.038 g/mL), 4-7 (1.046 - 1.079 g/mL), and 9-12 (1.106 - 1.259 g/mL). We performed t tests comparing the transformed expression levels over each RNA between each group of fractions to select informative ncRNAs. We selected those RNAs that showed significant difference (p value = 0.005) in the comparison against the rest of the reference samples. The final set contained 80 informative ncRNAs.

Stability Criterion

Each Stage 1 deconvolution requires the number of constituent cargo profiles to be provided for the model. In order to select the appropriate number of profiles, k, we generated 3 datasets using random 80% of the samples. Deconvolution was then performed with the number of constituent cargo types varying from 3 to 6. We compared the estimated profiles and proportions across the overlap in samples between each of the 3 subsets. The model that provided the highest correlations was selected. This process is fully automated in the EDec (Onuchic et al., 2016) R package.

Deconvolution Stage 1 (Estimate Constituent Cargo Profiles of Complex Biofluids)

Deconvolution is modeled after Onuchic et al. (2016). Instead of using methylation beta values for the deconvolution of constituent cell types within complex tissues, we used 0-1 transformed abundances, as described above.

Deconvolution Stage 2 (Estimate ncRNA Read Abundance of Constituent Cargo Profiles)

Stage 2 deconvolution is performed using the Read Counts or RPM sample profiles from the exRNA Atlas and the per-proportions estimated in Stage 1. To calculate the abundance of each ncRNA biotype, we multiply the per-sample proportion of each constituent cargo profile by the sum of all reads per biotype per constituent cargo profile. Abundances can be normalized to 100% by taking the ratio of each biotype over the total reads.

Identifying Differential Proportions of CTs between Case/Controls

A Wilcoxon Sign Test is used to compare the distribution of per-sample proportions of each constituent cargo profile between sample cohorts. Significance is indicated when p < 0.05.

Differential Expression of miRNAs between Case/Controls in Exercise Study

Stage 2 was run on Baseline and post-Exercise samples separately in order to predict the constituent cargo profiles in read counts for each cohort separately. Stage 2 outputs the mean expression and standard deviation for each miRNA. Using the mean and standard deviation, we performed a simple t test on each constituent cargo profile (Baseline versus Exercise) to identify differentially expressed miRNAs. Only miRNAs with a mean expression greater than 1 in all samples were tested. FDR correction (Benjamini & Hochberg) was used to account for multiple testing. miRNAs appear in their respective Venn as significant if FDR < 0.05 and log2(Fold Change) > 1 (Figure 6C, P1(CT2) circle, P2(CT3B) circle, P3(CT4) circle. For the Exercise Study, we also performed DESeq2 analysis to identify differentially expressed miRNAs between Baseline and post-Exercise cohorts (Figure 6C, *Shah et al.*, circle). This was done in order to adjust for exceRpt processing compared to the original publication. DESeq2 was run using the read counts for 2 groups: Baseline (n = 26) and post-Exercise (n = 26). Significant miRNAs were indicated if FDR < 0.05 and log2(Fold Change) > 1.

Pathway Analysis

miRNApath

We utilized the miRNApath: Pathway Enrichment for miRNA Expression Data (Cogswell et al., 2008) R package available on Bioconductor to identify miRNA-Gene-Pathway enrichment. miRNA-Gene association tables were downloaded from miRTarBase (interactions other than weak were filtered out) and provided to miRNApath. Gene-Pathway association tables are provided by miRNApath. Enrichment was performed on 3 sets of miRNAs (CT2, CT3B, CT4) using the default settings (Composite = TRUE, Permutations = 0). Pathways are indicated as significant if p value < 0.05. Figure 6D only includes pathways with a significance greater than 0.01.

Pathway Finder

Pathway Finder is a tool developed by the WikiPathways team and is integrated within the exRNA Atlas for viewing miRNAs in the context of biological pathways. WikiPathways (Slenter et al., 2018) is an open, collaborative pathway curation platform that publishes its data in a format readable by both humans and machines. Pathway Finder takes a user-specified list of miRNAs as input, combines that with pre-processed miRNA-to-gene mapping data, and produces as output a list of pathways that contain one or more miRNAs from the user-specified list and their gene targets. The results are displayed in a table format, with one pathway per row. Selecting a pathway from the table brings up an interactive diagram view of the result from WikiPathways with each miRNA and/or gene target highlighted. Recent upgrades to the tool have improved performance and usability.

STRING

Utilizing STRING, a database of protein-protein interaction networks, we calculated the enrichment of our predicted protein networks within each fraction using the whole biofluid as background knowledge (Szklarczyk et al., 2015).

DATA AND SOFTWARE AVAILABILITY

All exRNA Atlas data are publicly available at https://exrna-atlas.org. Atlas dataset accessions include: EXR-KJENS1WBaSro-AN (n = 523, GEO: GSE121868), EXR-KJENS1RID1-AN (n = 428, dbGaP: phs001258.v1.p1), EXR-KJENS1sPlvS2-AN (n = 345, dbGaP: phs000727.v1.p1), EXR-DWONG1qf3tcS-AN (n = 198, dbGaP: phs001767.v1.p1; GEO: GSE121870), EXR-MTEWA1cHYLo6-AN (n = 197, GEO: GSE121869), EXR-TPATE1OqELFf-AN (n = 192, GEO: GSE71008), EXR-MBITZ12SHVIr-AN (n = 80, GEO: GSE121978), EXR-KJENS12WGutU-AN (n = 70, GEO: GSE121867), EXR-SADAS1EXER1-AN (n = 62, GEO: GSE121874), EXR-SADAS1UJ0CzW-AN (n = 43, GEO: GSE121875), and EXR-KVICK1oIp40e-AN (n = 6, GEO: GSE121865). Note that GEO contains only processed data for EXR-KJENS1WBaSro-AN, EXR-DWONG1qf3tcS-AN, EXR-MTEWA1cHYLo6-AN, EXR-KJENS12WGutU-AN, EXR-SADAS1UJ0CzW-AN, and EXR-SADAS1UJ0CzW-AN - raw data for these datasets are either currently being processed for dbGaP or are undergoing IRB evaluation. The deconvolution algorithm can be found online: https://github.com/BRL-BCM/XDec. The deconvolution results for the 21 datasets used in this paper's deconvolution analysis are available via the Public Analysis Results page. The Genboree source code is distributed under a GNU Affero GPL v3.0 license and is available at https://github.com/BRL-BCM.

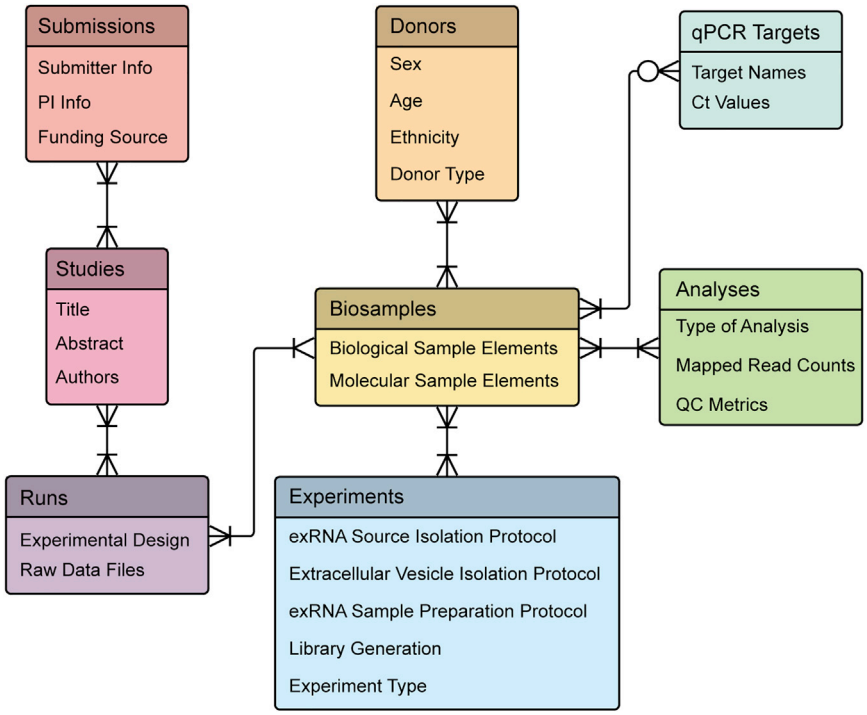


Figure S1. Overview of Metadata Stored in exRNA Atlas, Related to Figure 2

Infographic describing the key metadata entities modeled and stored in the exRNA Atlas. Relationships between different types of metadata are represented via connective lines. For instance, a given biosample may have zero, one, or many qPCR targets, depending on whether that biosample has been profiled using qPCR or RNA-seq (or both). In addition, donors may contribute multiple biosamples, and biosamples may come from multiple donors (via pooling).

Figure S2. exRNA Atlas Shows Large Amounts of Variability, Related to Figure 2

tSNE plot of exRNA Atlas samples show large amounts of variability across different studies.

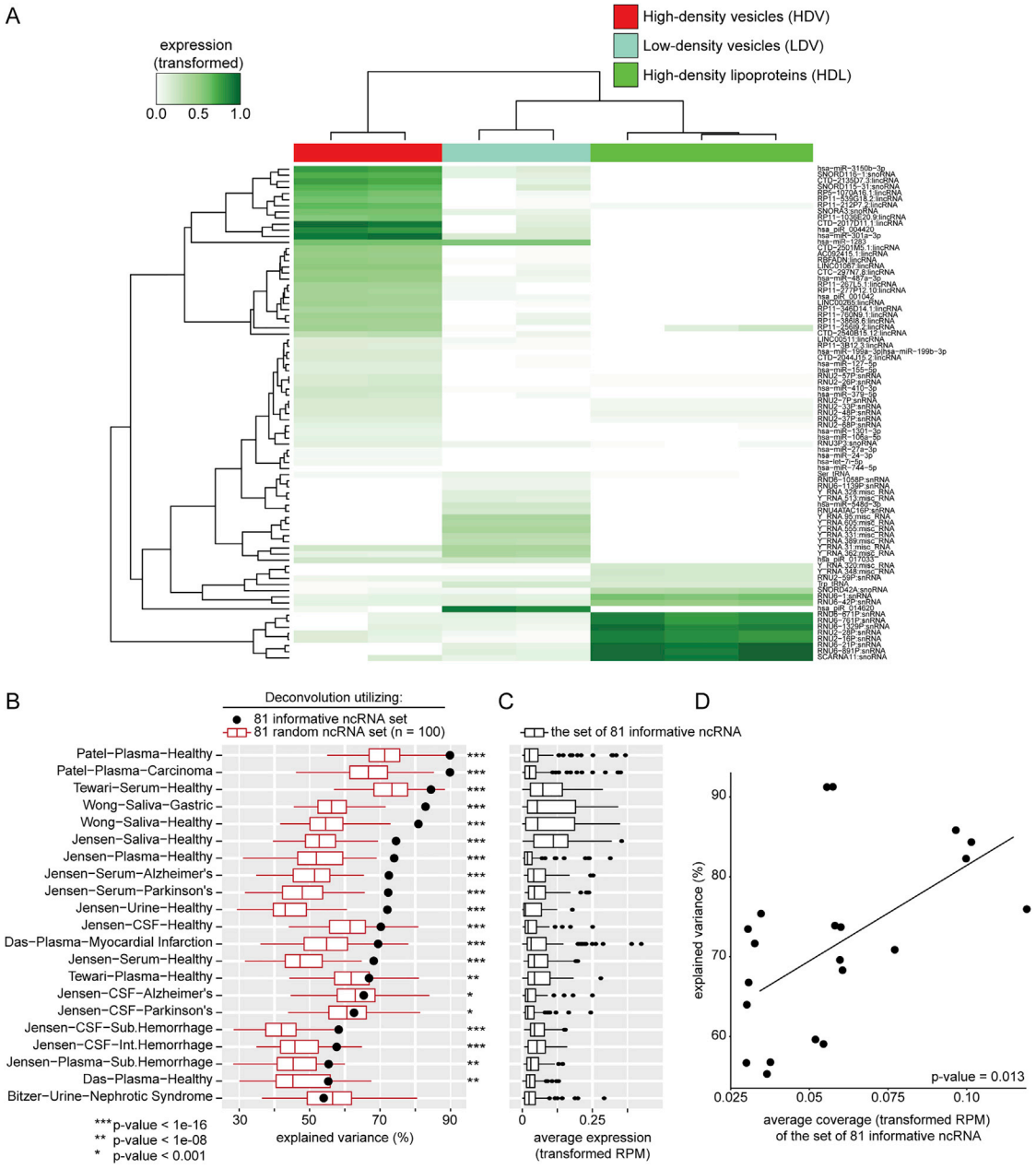


Figure S3. Selected Informative ncRNA Set Explains Greater Variance Than Random Sets, Related to Figure 3

(A) Heatmap representing the gene expression (normalized) pattern of 81 informative ncRNAs that are differentially expressed between the LD vesicles, HD vesicles, and HDL exRNA profiles (STAR Methods).

(B) We calculated the amount of explained variance when the deconvolution was utilized the 81 informative ncRNA set (Black Dot) that are differentially expressed between the HD, LD, and HDL exRNA profiles (Figure S3). Additionally, we generated 100 sets of random ncRNAs that contained the same number of RNAs per biotype as the informative set and measured the explained variance (Boxplot). p values were calculated using a one-sample t test.

(C) Boxplot represents the distribution of average expression of the 81 informative ncRNAs for each of the 21 analysis datasets.

(D) Each point indicates the correlation between the average expression of all 81 informative ncRNAs across all samples within each analysis dataset to the calculated explained variance when deconvolution is performed using the 81 informative ncRNA set.

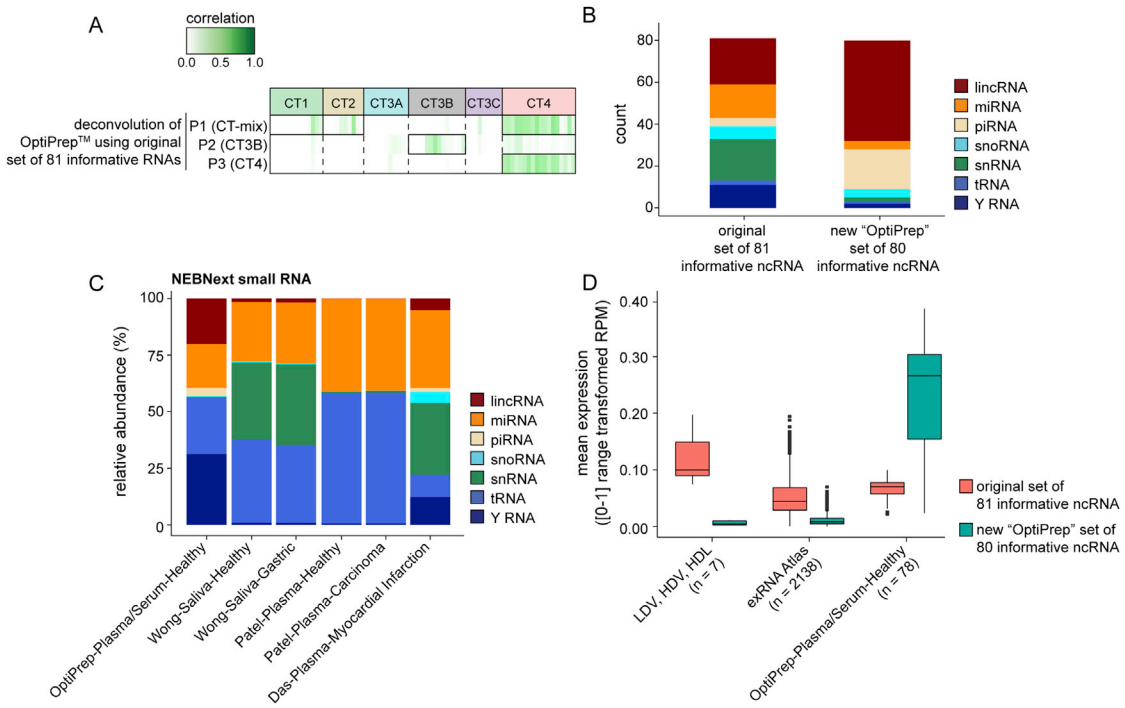


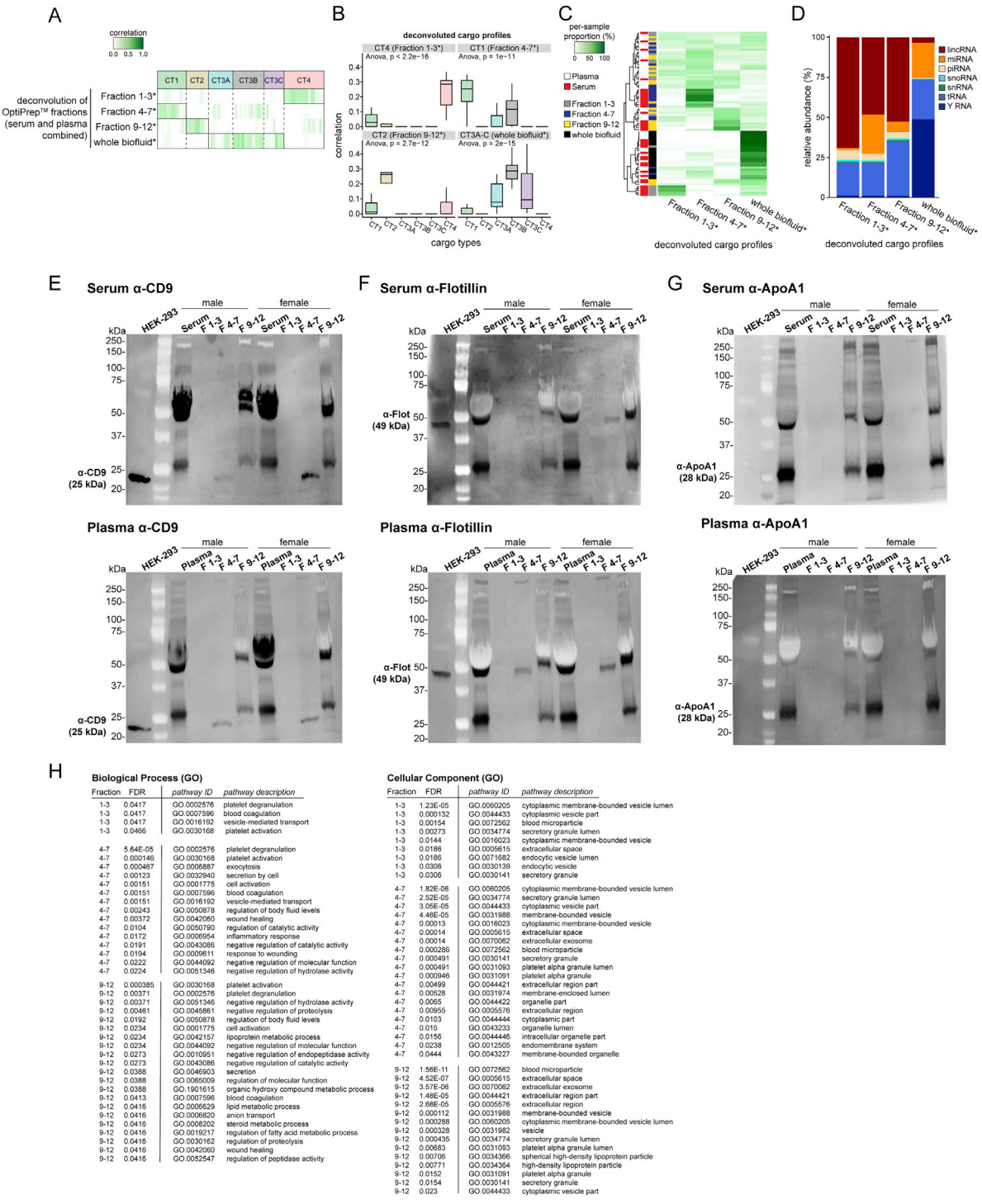
Figure S4. Comparison of Informative ncRNA Used for Deconvolution, Related to Figure 4

(A) Deconvolution was applied to 78 small RNA-seq profiles of the OptiPrep fractions and whole serum and plasma profiles using the original 81 informative ncRNAs. The algorithm estimated three cargo profiles. Heatmap correlation scores indicate that P1 is a mixture of CT1, CT2 and CT4. P2 and P3 correlate to a single cargo type CT3B and CT4, respectively.

(B) RNA count and biotype distribution of ncRNAs for the original set of 81 informative ncRNAs and the new "OptiPrep" set of 80 informative ncRNAs.

(C) Relative abundance of each ncRNA biotype for all datasets utilizing the NEBNext small RNA library preparation kit.

(D) Mean expression ([0-1] range transformation RPM) of the original set of 81 informative ncRNAs and the new "OptiPrep" set of 80 informative ncRNAs across the LDV, HDV, HDL samples, all exRNA Atlas samples, and the 78 OptiPrep fractions and whole biofluid samples.



Biological Process (GO)

Fraction	FDR	pathway ID	pathway description
1-3	0.0417	GO.0002576	platelet degranulation
1-3	0.0417	GO.0007596	blood coagulation
1-3	0.0417	GO.0016192	vesicle-mediated transport
1-3	0.0466	GO.0030168	platelet activation
4-7	5.64E-05	GO.0002576	platelet degranulation
4-7	0.000146	GO.0030168	platelet activation
4-7	0.000467	GO.0006887	exocytosis
4-7	0.00123	GO.0032940	secretion by cell
4-7	0.00151	GO.0001775	cell activation
4-7	0.00151	GO.0007596	blood coagulation
4-7	0.00151	GO.0016192	vesicle-mediated transport
4-7	0.00243	GO.0050878	regulation of body fluid levels
4-7	0.00372	GO.0042060	wound healing
4-7	0.0104	GO.0050790	regulation of catalytic activity
4-7	0.0172	GO.0006954	inflammatory response
4-7	0.0191	GO.0043086	negative regulation of catalytic activity
4-7	0.0194	GO.0009611	response to wounding
4-7	0.0222	GO.0044092	negative regulation of molecular function
4-7	0.0224	GO.0051346	negative regulation of hydrolase activity
9-12	0.000385	GO.0030168	platelet activation
9-12	0.00371	GO.0002576	platelet degranulation
9-12	0.00371	GO.0051346	negative regulation of hydrolase activity
9-12	0.00461	GO.0045861	negative regulation of proteolysis
9-12	0.0192	GO.0050878	regulation of body fluid levels
9-12	0.0234	GO.0001775	cell activation
9-12	0.0234	GO.0042157	lipoprotein metabolic process
9-12	0.0234	GO.0044092	negative regulation of molecular function
9-12	0.0273	GO.0010951	negative regulation of endopeptidase activity
9-12	0.0273	GO.0043086	negative regulation of catalytic activity
9-12	0.0388	GO.0046903	secretion
9-12	0.0388	GO.0065009	regulation of molecular function
9-12	0.0388	GO.1901615	organic hydroxy compound metabolic process
9-12	0.0413	GO.0007596	blood coagulation
9-12	0.0416	GO.0006629	lipid metabolic process
9-12	0.0416	GO.0006820	anion transport
9-12	0.0416	GO.0008202	steroid metabolic process
9-12	0.0416	GO.0019217	regulation of fatty acid metabolic process
9-12	0.0416	GO.0030162	regulation of proteolysis
9-12	0.0416	GO.0042060	wound healing
9-12	0.0416	GO.0052547	regulation of peptidase activity

Cellular Component (GO)

Fraction	FDR	pathway ID	pathway description
1-3	1.23E-05	GO.0060205	cytoplasmic membrane-bounded vesicle lumen
1-3	0.000132	GO.0044433	cytoplasmic vesicle part
1-3	0.00154	GO.0072562	blood microparticle
1-3	0.00273	GO.0034774	secretory granule lumen
1-3	0.0144	GO.0016023	cytoplasmic membrane-bounded vesicle
1-3	0.0186	GO.0005615	extracellular space
1-3	0.0186	GO.0071682	endocytic vesicle lumen
1-3	0.0306	GO.0030139	endocytic vesicle
1-3	0.0306	GO.0030141	secretory granule
4-7	1.82E-06	GO.0060205	cytoplasmic membrane-bounded vesicle lumen
4-7	2.52E-05	GO.0034774	secretory granule lumen
4-7	3.05E-05	GO.0044433	cytoplasmic vesicle part
4-7	4.48E-05	GO.0031988	membrane-bounded vesicle
4-7	0.00013	GO.0016023	cytoplasmic membrane-bounded vesicle
4-7	0.00014	GO.0005615	extracellular space
4-7	0.00014	GO.0070062	extracellular exosome
4-7	0.000286	GO.0072562	blood microparticle
4-7	0.000491	GO.0030141	secretory granule
4-7	0.000491	GO.0031093	platelet alpha granule lumen
4-7	0.000946	GO.0031091	platelet alpha granule
4-7	0.00499	GO.0044421	extracellular region part
4-7	0.00528	GO.0031974	membrane-enclosed lumen
4-7	0.0065	GO.0044422	organelle part
4-7	0.00955	GO.0005576	extracellular region
4-7	0.0103	GO.0044444	cytoplasmic part
4-7	0.015	GO.0043233	organelle lumen
4-7	0.0156	GO.0044446	intracellular organelle part
4-7	0.0238	GO.0012505	endomembrane system
4-7	0.0444	GO.0043227	membrane-bounded organelle
9-12	1.56E-11	GO.0072562	blood microparticle
9-12	4.52E-07	GO.0005615	extracellular space
9-12	3.57E-06	GO.0070062	extracellular exosome
9-12	1.48E-05	GO.0044421	extracellular region part
9-12	2.68E-05	GO.0005576	extracellular region
9-12	0.000112	GO.0031988	membrane-bounded vesicle
9-12	0.000288	GO.0060205	cytoplasmic membrane-bounded vesicle lumen
9-12	0.000328	GO.0031982	vesicle
9-12	0.000435	GO.0034774	secretory granule lumen
9-12	0.00683	GO.0031093	platelet alpha granule lumen
9-12	0.00706	GO.0034366	spherical high-density lipoprotein particle
9-12	0.00771	GO.0034364	high-density lipoprotein particle
9-12	0.0152	GO.0031091	platelet alpha granule
9-12	0.0154	GO.0030141	secretory granule
9-12	0.023	GO.0044433	cytoplasmic vesicle part

Figure S5. Deconvolution of Independent Fractionation Profiles, Related to Figure 4

(A) Deconvolution was applied to 78 small RNA-seq profiles of the OptiPrep fractions and whole serum and plasma profiles. The algorithm estimated four cargo profiles. Heatmap of correlation scores indicates that each of the four constituent profiles correlate to a given CT or the super group CT3.

(B) Boxplot of correlation scores between each estimated cargo profiles versus the 6 Cts. Based on the correlation score distributions, each cargo profile has significantly higher correlations to one CT (CT1, CT2, CT4) or group (CT3A-C) (ANOVA).

(C) Heatmap indicates the proportion of each predicted cargo profile (columns) within each sample (rows). Each cargo profile is named based on which fraction pool contains the highest proportions of that cargo profile.

(legend continued on next page)

(D) Estimated fraction of ncRNA contributed from each biotype to each deconvoluted profile based on the sum of all estimated Reads Per Million (RPM) for each ncRNA gene within each RNA biotype. Relative abundance was calculated by dividing by the sum of all reads for those 7 RNA biotypes.
(E) western blot analysis of anti-CD9 protein marker for serum and plasma: HEK293 (control), male and female: whole biofluid, fractions 1-3, fraction 4-7, fraction 9-12. Expected size for CD9 is 25 kDA.
(F) western blot analysis of anti-Flotillin protein marker for serum and plasma: HEK293 (control), male and female: whole biofluid, fractions 1-3, fraction 4-7, fraction 9-12. Expected size for Flotillin is 49 kDA.
(G) western blot analysis of anti-ApoA1 protein marker for serum and plasma: HEK293 (control), for male and female: whole biofluid, fractions 1-3, fraction 4-7, fraction 9-12. Expected size for ApoA1 is 28 kDA.
(H) Mass Spectrometry was performed on each of the OptiPrep pooled fractions (Fraction 1-3, Fraction 4-7 and Fraction 9-12) across all 5 males and 5 females. Protein counts were summed for each fraction. Pathway enrichment proteins detected by STRING results indicate both Biological Process enrichment and Cellular Component enrichment.

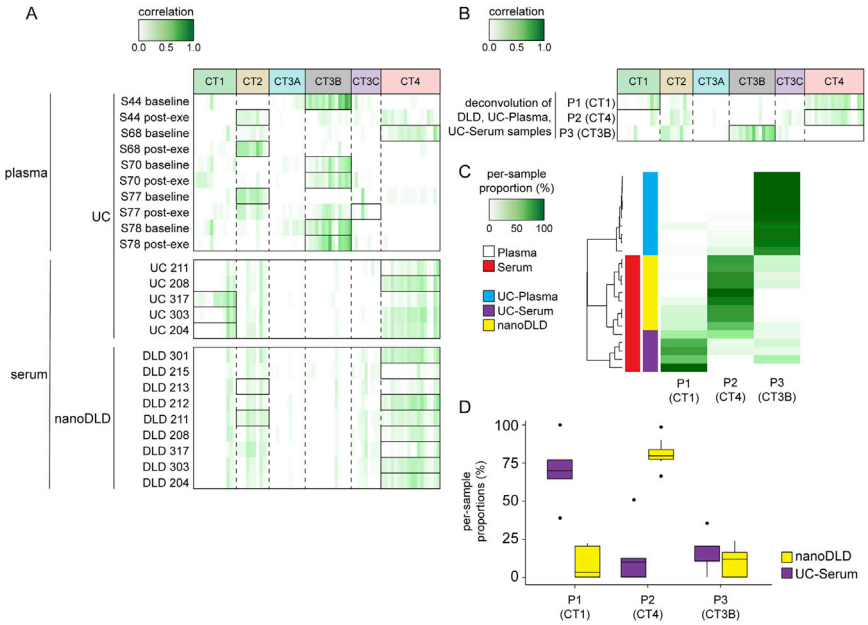


Figure S6. Deconvolution of Extracellular Vesicles and Isolated NanoDLD Extracellular Vesicles, Related to Figure 4

(A) Heatmap represents the correlation scores between the 75 cargo profiles and the profiles of ten extracellular vesicle profiles isolated from healthy plasma samples deposited in the Atlas. Serum-Ultracentrifugation heatmaps represent the correlation scores between the 75 predicted cargo profiles and the profiles (n = 5) of extracellular vesicle profiles isolated from prostate cancer serum samples. Serum nanoDLD heatmap represent the correlation scores between the 75 cargo profiles and the profiles (n = 9) of extracellular vesicle profiles isolated from prostate cancer serum samples using nano-DLD technology. All correlations are estimated across the 81 informative ncRNAs.

(B) Deconvolution was applied to 24 small RNA-seq profiles of UC-Plasma (n = 10), UC-Serum (n = 5) and nanoDLD extracellular vesicle (n = 9) exRNA profiles. The algorithm estimated three cargo profiles. Heatmap of correlation scores indicates that each of the three profiles correlate to a given CT.

(C) Heatmap indicates the proportion of each predicted cargo profile (columns) within each sample (rows).

(D) Box-plot indicating the per-sample proportions of each of the predicted cargo profiles specifically for the UC-Serum and nanoDLD samples.

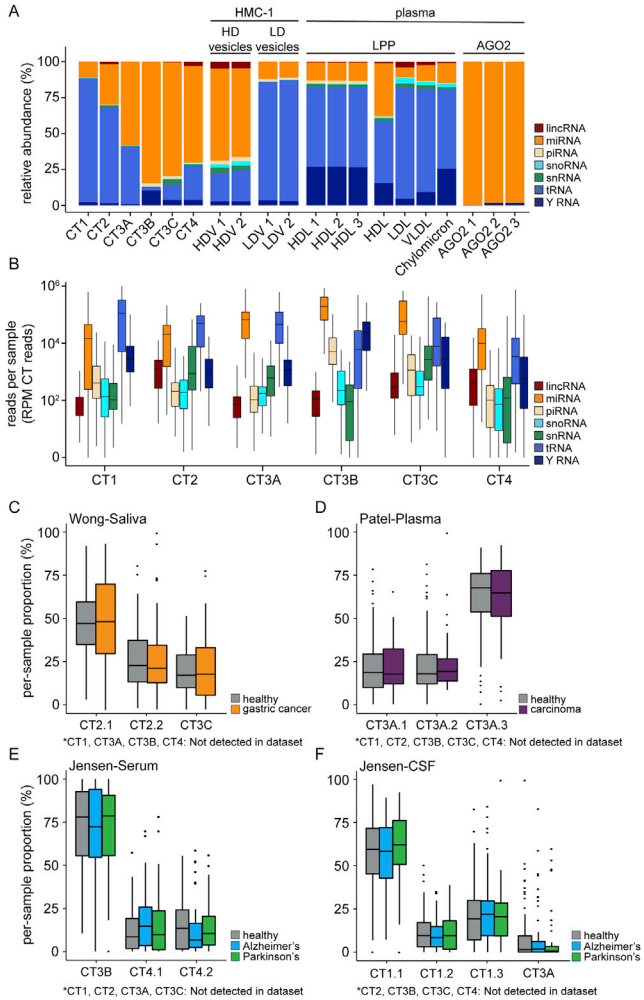


Figure S7. Abundance of ncRNA Biotypes to Each Cargo Type and Per-Sample Proportions of Constituent Cargo Profiles within Case-Control Studies, Related to Figures 4 and 7

(A) Bar chart of the median relative abundances across all deconvoluted members clustered within each CT for each of the predicted CT1-4. Additionally, we included the profiles of physically isolated exRNA profiles (HD vesicles, LD vesicles, LPP, AGO2). We estimated the fraction of ncRNA species in each cargo type by summing the reads of each ncRNA biotype and dividing by the sum of all estimated ncRNA Reads (displayed as Reads Per Million (RPM)).

(B) Boxplot indicating the distribution of abundance of reads per ncRNA biotype per sample. Median values were used to estimate the relative abundance in (A).

(C) Per-sample proportions of Wong Saliva samples show no differential abundance between healthy and samples from gastric patients. Wilcoxon Sign Test was used to determine if proportion distribution was significantly different.

(D) Per-sample proportions of Patel Wong samples show no differential abundance between healthy and samples from gastric patients. Wilcoxon Sign Test was used to determine if proportion distribution was significantly different.

(E) Per-sample proportions of Jensen Serum samples show no differential abundance between healthy and samples from gastric patients. Wilcoxon Sign Test was used to determine if proportion distribution was significantly different.

(F) Per-sample proportions of Jensen CSF samples show no differential abundance between healthy and samples from gastric patients. Wilcoxon Sign Test was used to determine if proportion distribution was significantly different.