SCIENTIFIC REPORTS

OPEN

A total transcriptome profiling method for plasma-derived extracellular vesicles: applications for liquid biopsies

Received: 22 May 2017
Accepted: 9 October 2017
Published online: 31 October 2017

Maria G. Amorim[1], Renan Valieris[2], Rodrigo D. Drummond[2], Melissa P. Pizzi[1], Vanessa M. Freitas[3], Rita Sinigaglia-Coimbra[4], George A. Calin[5], Renata Pasqualini[6], Wadih Arap[6], Israel T. Silva[2,7], Emmanuel Dias-Neto[1,8] & Diana N. Nunes[1]

Extracellular vesicles (EVs) are key mediators of intercellular communication. Part of their biological effects can be attributed to the transfer of cargos of diverse types of RNAs, which are promising diagnostic and prognostic biomarkers. EVs found in human biofluids are a valuable source for the development of minimally invasive assays. However, the total transcriptional landscape of EVs is still largely unknown. Here we develop a new method for total transcriptome profiling of plasma-derived EVs by next generation sequencing (NGS) from limited quantities of patient-derived clinical samples, which enables the unbiased characterization of the complete RNA cargo, including both small- and long-RNAs, in a single library preparation step. This approach was applied to RNA extracted from EVs isolated by ultracentrifugation from the plasma of five healthy volunteers. Among the most abundant RNAs identified we found small RNAs such as tRNAs, miRNAs and miscellaneous RNAs, which have largely unknown functions. We also identified protein-coding and long noncoding transcripts, as well as circular RNA species that were also experimentally validated. This method enables, for the first time, the full spectrum of transcriptome data to be obtained from minute patient-derived samples, and will therefore potentially allow the identification of cell-to-cell communication mechanisms and biomarkers.

"Liquid biopsies" are being increasingly recognized as transformative in biology and medicine. Within such context, extracellular vesicles (EVs) such as exosomes and microvesicles are involved in a wide variety of physiological processes and have important roles in cell-to-cell communication during development, as well as in health and diseased states[1,2]. Their capacity to influence the physiology of the recipient cells/tissues is due to the transfer of their cargo of lipids, proteins, and nucleic acids[3,4], which is produced by their parental cells, selected and loaded into the EVs[5], and delivered both locally and to distant sites[6,7]. In this sense, the characterization of the full repertoire of EVs-cargo is not only relevant for understanding their potential biological roles, but can also be seen as a source of potential biomarkers of diagnostic and prognostic value in the setting of a wide range of pathological conditions, including cancer, autoimmune or inflammatory, as well as and neurological and infectious diseases.

The key for determining EV content is recovering sufficient amounts of vesicles from patient samples. This challenge is particularly evident in the characterization of EVs present in the peripheral blood of patients, where

[1]Laboratory of Medical Genomics, A.C.Camargo Cancer Center, São Paulo, SP, Brazil. [2]Laboratory of Computational Biology, A.C.Camargo Cancer Center, São Paulo, SP, Brazil. [3]Department of Cell and Developmental Biology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, SP, Brazil. [4]Electron Microscopy Center, Federal University of São Paulo, São Paulo, SP, Brazil. [5]Department of Experimental Therapeutics and Center for RNA Interference & Non Coding RNAs, The University of Texas MD Anderson Cancer Center, Houston, USA. [6]University of New Mexico Comprehensive Cancer Center and Divisions of Hematology/Oncology and Molecular Medicine, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM, USA. [7]Laboratory of Molecular Immunology, The Rockefeller University, New York, NY, USA. [8]Laboratório de Neurociências Alzira Denise Hertzog Silva (LIM27), Instituto de Psiquiatria, FMUSP, São Paulo, SP, Brazil. Correspondence and requests for materials should be addressed to E.D.-N. (email: emmanuel@cipe.accamargo.org.br) or D.N.N. (email: dnoronha@cipe.accamargo.org.br)

often only a few milliliters of blood might be available for research investigation, especially in patients with poor clinical conditions and/or advanced disease. Thus far, this practical limitation has hindered a comprehensive analysis of vesicular cargo, and thereby prevented the exploration of the full potential of EVs for clinical applications.

RNA molecules, including microRNAs, long noncoding RNAs and viral RNAs, carried by EVs are amongst the most promising biomarkers for the detection and monitoring of disease[3,8,9], and may perhaps also be used for monitoring therapeutic response. Notably, recent studies have attempted to profile populations of vesicular RNAs by using next generation sequencing (NGS), to allow the identification of a catalogue of vesicle-derived RNAs (Table 1**)**. However, most of these studies employed size-selection protocols during NGS library preparation, which has limited the analysis essentially to small RNAs[10–13]. On the other hand, a recent report, has only analyzed RNAs larger than 50 nt, which has essentially excluded molecules such as mature miRNAs[14]. Similarly, amplification steps with oligo-dT primers are also restricted to the study of the polyadenylated fraction of the transcriptome[15].

In order to address such technical limitations, here we have developed a streamlined new methodology of total transcriptome profiling of EVs by NGS, which includes the simultaneous analysis of both small and long RNAs. In order to do so, total RNAs were fragmented by enzymatic digestion and technical modifications for library construction were adopted, leading to an increase in the total number of molecules, which subsequently led to reduced adaptor-dimers. Moreover, by not excluding RNAs of any particular size, our method allows an unbiased characterization of the total transcriptional landscape of EVs, which still remains largely unknown.

Material and Methods

Blood collection, EV isolation, and RNA extraction. This study was approved by the research ethics committee of A.C. Camargo Cancer Center (ACCCC; protocol 1554/11). Blood samples were collected in BD Vacutainer tubes with Acid Citrate Dextrose (ACD) solution, from healthy volunteer individuals participating in the cancer prevention campaign at the ACCCC (specifically, a total of five women with no mammogram evidence of breast cancer), after signing an Institutional Review Board-approved written informed consent form. Plasma samples were separated after two centrifugations at 2,500 × g for 15 min and stored at −80 °C. EVs were isolated from 2.5 ml of plasma by ultracentrifugation, as described[19], but no filtering steps were performed, to allow a broader view of the circulating EVs transcriptome. After the centrifugations at 2,000 × g for 30 min and 12,000 × g for 45 min (Eppendorf centrifuge 5810R, fixed-angle rotor F-34-6-38), pellets were discarded to eliminate cellular debris and apoptotic bodies. Supernatants were then centrifuged at 110,000 × g for 120 min and the resulting EV-containing pellets were washed once in phosphate-buffered saline (PBS) and centrifuged again at 110,000 × g for 70 min (Beckman ultracentrifuge Optima L-90K, swinging rotor SW-41Ti, polypropylene tubes Beckman 331372). PBS was discarded and the final EV-pellet was finally resuspended in 350 µl of lysis solution and RNA was extracted by using spin columns (Total RNA Purification Kit – Norgen Biotek, Canada). As large quantities of RNA were needed for optimization of the RNA fragmentation step, RNA derived from blood buffy coat was used for fragmentation-standardization purposes. For this purpose, blood was collected in two Vacutainer tubes (Becton Dickinson, USA) containing di-potassium ethylenediaminetetraacetic acid (K_2 EDTA); plasma was separated after two centrifugations at 2,500 × g for 15 min, the buffy coat (100 µl) was transferred to a fresh Eppendorf tube followed by the addition of 350 µl of lysis solution and RNAs were extracted (total of four separate extractions) by using spin columns (Total RNA Purification Kit – Norgen Biotek, Canada).

Electron microscopy. EVs isolated from 4.5 ml of plasma (as described above) and resuspended in at least 50 µl PBS after ultracentrifugation, were washed, pelleted again and finally resuspended in 50 µl of PBS containing 4% methanol-free formaldehyde. Whole-mount negative staining of EVs was performed as described[19]. Briefly, a nickel formvar/carbon-coated grid (Electron Microscopy Sciences, USA) was floated on top of a drop of 15 µl fixed-EVs sample for 20 min, and then washed three times with sodium cacodylate buffer, fixed with PBS containing 1% glutaraldehyde for 5 min, washed seven times with double-distilled water, contrasted with uranyl oxalate for 5 min, embedded with a admixture of ice-cold uranyl acetate/methyl cellulose (1:9 v/v) for 10 min, then dried and analyzed in a JEM-1010 electron microscope (JEOL, Tokyo, Japan) at 80 kV.

NanoSight. EVs isolated from 200 µl of plasma (as described above) were resuspended in 500 µl of filtered PBS and evaluated by a NanoSight LM10 (Malvern, UK). Images were acquired for 60 sec (triplicates for each sample) with the following parameters: camera shutter — 1495; camera gain — 512; detection threshold — 10.

Western blot. For protein extraction, EV-pellets (obtained after ultracentrifugation of a pool of four plasma samples, total 6 ml) were resuspended in 100 µl lysis buffer containing 300 mM NaCl, 50 mM Tris pH 7.4, 0.5% NP-40 and anti-proteases cocktail (Roche, Indianapolis, IN, USA). Protein concentrations were determined by the Bradford assay (Bio-Rad, Hercules, CA, USA), and 15 µg total protein was submitted to sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). Ponceau staining was performed before blocking with 5% milk. Primary commercially available antibodies used were CD63 1:1,000 (CBL553, Millipore, Billerica, MA, United States), RAB27B 1:250 (HPA019849, Sigma-Aldrich), Flotillin 1:500 (ab41927, Abcam, Cambridge, MA, USA), HSP70 1:1,000 (EXOAB-Hsp70A-1, System Biosciences, Mountain View, CA, USA), TSG101 1:500 (ab4A10, Abcam, Cambridge, MA, USA) and RAB7A 1:1,000 (ab50533, Abcam, Cambridge, MA, USA).

RNA fragmentation optimization. Each aliquot of total RNA, extracted from buffy-coat as described above, was treated with 2U DNAse (TURBO DNA-free kit – Thermo Fisher Scientific, USA), pooled, purified and concentrated with RNeasy MinElute Clean-up Kit – Qiagen, USA). RNA was quantified by NanoDrop and individual 500ng RNA aliquots (in 8 µl of RNase-free water) were fragmented at 37 °C with 1 µl RNAse III and 1 µl 10X RNAse III Reaction Buffer (both from Thermo Fisher, USA), for the following time points: 15, 30, 45, 75, 105, 135, 180, 240, 300, 360 and 420 min. After each time point, 20 µl of RNase-free water were added to each

Reference	Sample type	Number of EVs samples	EVs isolation methodology	rRNA depletion	Input	RNA-Seq methodology	Raw reads (sample)	Mapped reads (sample)	rRNA	# Molecules identified	Major categories of molecules identified	Technical validation
Nolte-'t Hoen *et al.*, 2012[11]	CCM from mouse DC-T cell co-culture	1	Ultracentrifugation	No	450 ml CCM	Small RNA (15–70 nt) size-selection on native 6% gradient PAGE gel Sequencing on Illumina (single-end 35 cycles)	28.8 M	27.5 M	~30%	NA	tRNA repeat, Simple repeat LINE, rRNA, vRNA, Protein coding, SRP-RNA, Y-RNA	qPCR for miR-29a, miR-155, miR-191, Y-RNA, SRP-RNA
Huang *et al.*[10]	Human plasma	3 = 14 libraries	ExoQuick	No	250 µl plasma/2ng RNA	Small RNA (20–40 nt) size-selection on native 5% acrylamide gel Sequencing on Illumina HiSeq. 2000 (single-end)	7.27 M	4.04 M	9.16%	593 miRNAs	miRNA (76.20% of all mappable reads), ribosomal RNA (9.16%), long noncoding RNA (3.36%), piwi-interacting RNA (1.31%), transfer RNA (1.24%), small nuclear RNA (0.18%), and small nucleolar RNA (0.01%); fragments of coding sequence (1.36%), 5′ untranslated region (0.21%), and 3′ untranslated region (0.54%)	qPCR for miR-92a-3p, miR-191-3p, miR-26b-5p
Jenjaroenpun *et al.*[16]	CCM from MDA-MB-231 and MDA-MB-436, human metastatic breast cancer cell lines	2	Ultracentrifugation	RiboMinus (Thermo) did not work, authors argue due to fragmented rRNA	48 h CCM/ 200ng RNA	Whole transcriptome library with RNAse III fragmentation Considered only reads ≥ 20 bp Sequencing on Ion Torrent PGM	3.3 M	3 M	97%	16,086 transcripts (RPKM ≥ 1)	small nucleolar RNA, small nuclear RNA, Mt_tRNA, microRNA	qPCR for GAPDH, EEF1A1, FTH1, FTL, RAB13, RPPH1, RPL28
Schageman *et al.*[17]	CCM from HeLa cells and human serum	3 = 7 libraries	Total exosome RNA and protein isolation kit (Thermo Fisher)	No	4 ml serum/~2ng RNA	Small RNA bead-based Sequencing on Ion Torrent PGM Considered only reads > 17 bp	5–6 M	90–98%	~15–40%	NA	mRNA, rRNA, tRNA, miRNA, ncRNA	serum EVs - qPCR for has-mir-1281, has-mir-4257, has-mir-451
Li *et al.*[18]	Human serum and urine	4 = 8 libraries	Total exosome RNA and protein isolation kit (Thermo Fisher)	No	4 ml serum/10 ml urine/ ~2ng RNA	Small RNA bead-based Sequencing on Ion Torrent PGM	5–6 M	90–98%	serum: 5–30% urine: 30–60%	NA	miRNA, rRNA, tRNA, mRNA, piRNA	NA
Yuan *et al.*[12]	Human plasma	192	ExoQuick	No	2–10ng RNA	Small RNA (20–40 nt) size-selection on native 5% acrylamide gel Sequencing on Illumina HiSeq. 2000	12.6 M	5.4 M	0.7%	3,387 RNAs	miRNAs (~40.4%), piwiRNAs (~40.0%), pseudo-genes (~3.7%), lncRNAs (~2.4%), tRNAs (~2.1%), and mRNAs (~2.1%)	No
Lefebvre *et al.*[14]	CCM from A431 epidermoid carcinoma and HepG2 hepatocellular carcinoma cell lines	2	Ultracentrifugation	No	80 ml CCM/ 50ng RNA	No size selection was performed - only analysed RNAs > 50 nt Sequencing on Illumina HiSeq. 2000	NA	6.88 M	91.19%	7,361 RNAs (FPKM ≥ 5)	miscRNA, mRNA, lincRNA, snRNA, snoRNA, antisense	qPCR for TPT1, PABPC1, ATF4, PTBP1, HDGF, G3BP1,BRAF
San Lucas *et al.*[15]	Human pleural effusion and plasma	3	Ultracentrifugation	No	800 ml pleural effusion 15 ml plasma	mRNA amplification using oligo dT primers Sequencing on Illumina HiSeq. 2500	NA	498 M	NA	NA	Protein coding	No
Quek *et al.*[13]	CCM from GT1-7 mouse hypothalamic neuronal cell lines	5 = 23 libraries (4 gradient fractions, 1 UCexo)	Ultracentrifugation and Optiprep gradient	No	96 h CCM/ 20ng RNA	Small RNA bead-based Sequencing on Ion Torrent PGM	NA	1 M	0.54%	515 small RNAs	fragments of tRNA (range 41.6–67.0%), fragments of RNA repeat elements (~42.26%), miRNA (range 0.48–2.11%), protein-coding mRNA (range 0.49–1.82%), piRNA (range 0.46–1.71%), snRNA (range 0.11–0.23.%), snoRNA (range 0.04–0.10%), rRNA (range 0.20–0.88%)	qPCR and dPCR for let-7b and miR-342-3p

Table 1. Summary of the recent reports employing RNA sequencing analysis of EVs. M: million, nt: nucleotides, bp: base pairs, CCM: cell conditioned medium, EVs: extracellular vesicles, NA: information not available, RPKM: reads per kilobase per million mapped reads, FPKM: fragments per kilobase of transcript per million mapped reads, RPM: reads per million, qPCR: quantitative real time PCR, UCexo: exosomes from ultracentrifugation, dPCR: digital PCR.

RNA-aliquot, which was further purified with magnetic Nucleic Acid Binding Beads (Thermo Fisher, USA). The size-range of each aliquot of fragmented RNA was visualized with Bioanalyzer Pico and Small RNA chips (Agilent, USA).

Library construction and sequencing. Whole transcriptome libraries were constructed by using reagents provided in the Ion Total RNA-Seq Kit v2 (Thermo Fisher, USA), following the user guide of the Total Exosome RNA and Protein Isolation Kit (Thermo Fisher, USA), but including several protocol modifications, marked with an asterisk (*). Due to the limited yield, the whole eluate volume of 50 µl was vacuum concentrated to 16 µl, to allow all the extracted mass of RNA to be used. The total RNA volume was then fragmented by RNAse III at 37 °C for 180 min*, following the standardization described above. This step served to increase the number of RNA molecules available to hybridize with adaptors, minimizing the formation of adaptor dimers and thereby allowing the simultaneous analysis of the whole transcriptome, including all size-classes of RNAs, in a one-step library preparation protocol. The resulting RNA molecules were purified with magnetic Nucleic Acid Binding Beads* (following the recommendation in the Ion Total RNA-Seq Kit v2 protocol for Whole Transcriptome libraries, as opposed to spin columns, in order to minimize RNA loss), and hybridization to the Ion Adaptor Mix v2 was performed at 65 °C for 30 min* (the incubation time was extended to allow more time for the RNA molecules to hybridize to the adaptors), followed by ligation at 16 °C for 16 h. Templates were denatured at 75 °C for 15 min* (the incubation time and temperature were increased to assure a total denaturation of RNA secondary structures), followed by cDNA synthesis at 42 °C for 30 min. After bead-purification, the cDNAs were amplified and barcodes added to each sample by PCR* (an extra 2 PCR cycles were performed and extension steps were removed to disfavor the amplification of longer fragments). After bead-purification of amplified libraries, molarities of final libraries were determined by quantification on a Bioanalyzer instrument using the High Sensitivity DNA Kit (Agilent, USA). Libraries were diluted to the recommended concentration for emulsion PCR on OneTouch2 instrument (Thermo Fisher, USA). Enriched samples were finally deposited on semiconductor chips (P1) for sequencing on the Ion Torrent Proton platform.

Bioinformatics analysis. The bioinformatics pipeline started with the filtering of low quality reads (Torrent Suite 5.0 trims the reads to achieve >Phred 15 in a sliding-window of 30 bases). The annotation of the total transcriptome consisted of an initial step with the FastQ Screen[20] to remove rRNAs, repetitive, as well as other non-human DNA elements. miRDeep2 software[21] was used to annotate miRNAs in miRBase v20 and unmapped reads were aligned against hg19 genome reference, by using the STAR aligner[22]. tRNAs were annotated by using GtRNAdb database, followed by Ensembl Homo_sapiens.GRCh37.73 for all other transcript annotations. Reads with low mapping scores (Phred < 20, indicating alignment to multiple genomic loci) were removed. To filter out genes with non-uniform reads distribution along the exons, likely to be derived from spurious mapping, we adopted an approach based on Gini normalization coefficient[23] as a measure for the mRNA coverage variation (mcv). For this purpose, the mRNA for each RefSeq gene was divided into 10 sub-regions designated as 'bins'. In each such bin, we counted the number of aligned reads and applied to it the Gini coefficient to calculate the mcv-score. The genes with mcv-scores <0.7 were retained for further analysis. The circRNAs prediction was performed with the algorithm find_circ, following the default settings as described[24] (Fig. 1).

qRT-PCR validation. Quantitative RT-PCR (qRT-PCR) was performed to confirm the intra-vesicular origin of selected molecules, as well as the quantification inferred from NGS-data. For this purpose, EVs pellets derived from 0.5 ml of plasma, were isolated by ultracentrifugation as described above, resuspended in PBS and treated with RNase-A (Thermo Fisher, USA) in 0.5 µg/µl final concentration at 37 °C for 20 min. After treatment, RNase-A was inactivated with 1U/µl of RNase Inhibitor (New England Biolabs, USA) at room temperature (RT) and RNA was extracted as described above. To verify the RNase-A activity, a parallel preparation was performed with the same samples wherein the EVs pellet was resuspended in lysis buffer (100 mM Tris, 5 mM EDTA, 0.2% SDS, 0.2 M NaCl, 0.1 ml/ml proteinase K), followed by proteinase-K inactivation at 90 °C for 5 min and RNase-A incubation and treatment with RNase inhibitor and RNA extraction as above. cDNAs were synthesized with miScript II RT Kit (Qiagen, USA) in 20 µl reactions containing 12 µl of RNA, 4 µl of 5X miScript HiSpec Buffer, 2 µl of 10X miScript Nucleics Mix and 2 µl of miScript Reverse Transcriptase Mix, which was incubated at 37 °C for 60 min and 95 °C for 5 min. cDNAs were pre-amplified with miScript PreAMP PCR Kit (Qiagen, USA) in 25 µl reaction containing: 5 µl of cDNA diluted 1:5, 5 µl of miScript PreAMP Buffer, 2 µl of HotStarTaq DNA Polymerase, 5 µl of pool of miRNA assays of interest, 7 µl of nuclease-free water, and 1 µl of miScript PreAMP Universal Primer. Reactions were incubated at 95 °C for 15 min, followed by 12 cycles of 94 °C for 30 sec and 60 °C for 3 min. Pre-amplified cDNAs were analyzed in a 7500 Fast Real-Time PCR System (Thermo Fisher, USA) with miScript SYBR Green PCR Kit (Qiagen, USA) in 10 µl reaction containing 5 µl of 2X QuantiTect SYBR Green PCR Master Mix, 1 µl of 10X miScript Universal Primer, 1 µl of 10X miScript primer assay (miR-223-3p: MS00003871, let-7g-5p: MS00008337), 2 µl of nuclease-free water, and 1 µl of pre-amplified cDNA diluted 1:20.

circRNA validation. Validation of a representative circular RNA, transcribed from the *CORO1C* gene, was performed by PCR amplification by using outward primers, followed by Sanger sequencing (Suppl. Figure 1). For this purpose, after RNA extraction as described above, cDNA was synthesized by using the SuperScript III First-Strand Synthesis System for RT-PCR (Thermo Fisher, USA). The 20 µl reaction consisted of 8 µl RNA, 50ng of random hexamers, 1 µl of 10 mM dNTP mix, which was incubated at 65 °C for 5 min, 4 °C for 1 min, followed by the addition of the reagents: 2 µl 10X RT buffer, 4 µl 25 mM MgCl2, 2 µl 0.1 M DTT, 1 µl RNaseOUT (40U/µl) and 1 µl SuperScript III RT (200U/µl). The reactions were incubated at 25 °C for 10 min, 50 °C for 50 min, 85 °C for 5 min, 4 °C for 1 min. Finally, 1 µl of RNase H (2U/µl) was added to the reactions and incubated at 37 °C for 20 min.

The PCR primers were designed in such a way to amplify either the linear (inward facing primers, LF and LR) or circular isoforms (outward facing primers, CF and CR). The sequences (5′−3′) of primers used are:

LF: GCTGCTGAATGTGTTGAGGT and LR: GCTGCCTTTCTATGACCCTG; CF: ACCTCAACACAT TCAGCAGC and CR: CAGGGTCATAGAAAGGCAGC.

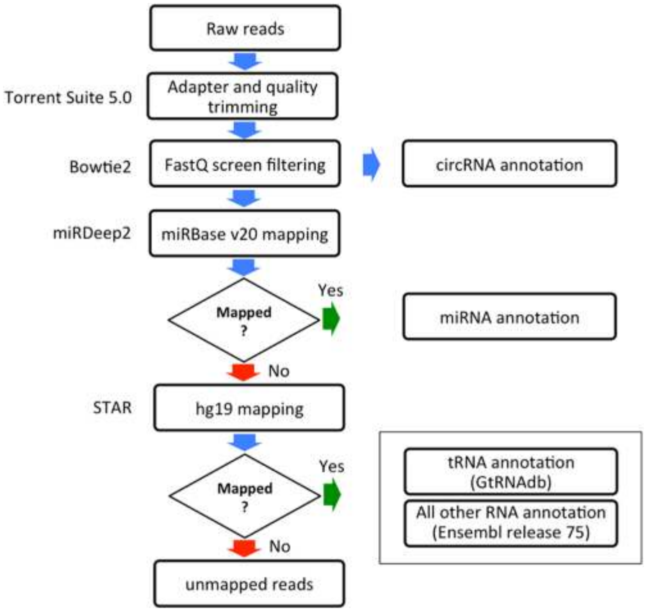

Figure 1. Scheme of the bioinformatics analysis pipeline.

The 15 µl PCR consisted of: 7.5 µl 2X GoTaq Green Master Mix (Promega, USA), 2 µl F primer at 2.5pmol/µl, 2 µl R primer at 2.5 pmol/µl, 2.5 µl nuclease- free water and 1 µl cDNA. PCR cycling conditions were: 95 °C for 2 min followed by 40 cycles at 95 °C for 1 min, 60 °C for 30 sec, 72 °C for 30 sec and final extension at 72 °C for 5 min. Afterwards 10 µl of each PCR product was evaluated on 8% polyacrylamide gel stained with silver[25] and 5 µl was purified with ExoSAP-IT (Affymetrix, USA) and sequenced by the Sanger method.

Data availability. All annotation of mapped reads generated or analysed during this study are included in this published article (and its Supplementary Information files). Raw data are available from the corresponding author on reasonable request.

Results

EVs characterization. We used blood samples isolated from healthy women (n = 5) and isolated EVs as described[19]. To certify that we were indeed working with *bona fide* EVs, the isolated particles were first validated by transmission electron microscopy (TEM), Nanoparticle Tracking Analysis (NTA) and Western blotting. TEM revealed the characteristic "cup-shaped" EVs with morphology, which has been previously described[19] and sizes typical of exosomes (Fig. 2A), and NTA showed the expected size-range (174–193 nm, and an average size of 185 nm - considering averages from the five plasma samples) an average concentration of 8 × 10^8 particles/ml (range 4–18 × 10^8 particles/ml), consistent with published reports[10,15]. Most EVs were smaller than 250 nm, consistent with a heterogeneous mixture of exosomes and microvesicles (Fig. 2B). Western blot analysis of EV-pellet obtained from a pool of plasma samples confirmed the expression of a panel (n = 6) of standard vesicular protein markers: CD63, FLOTILLIN, HSP70, RAB27B (Fig. 2C) and TSG101, RAB7A (Suppl. Figure 3).

Optimization of RNA input for NGS library-construction. Next, we developed a broad protocol to isolate the full spectrum of RNA cargos from the EV samples. First, we used RNA fragmentation to increase the number of molecules as a means of reducing the formation of adaptor-dimer artifacts, which consume sequencing space and it is a known major issue when limiting amounts of nucleic acids are available for NGS library construction[26]. As EV-derived RNA obtained from plasma cannot be generally visualized in the Bioanalyzer due to its very low amounts, the fragmentation time was optimized with RNA extracted from peripheral blood leukocytes, analyzed with the Agilent RNA 6000 Pico Kit. Results showed the expected 18S and 28S rRNA peaks in non-fragmented RNA (red traces), while RNA fragmented for 180 min (light green traces) consisted of fragments <200 nt in length (Fig. 3A). Analysis with Agilent Small RNA kit of further fragmentation time-points showed that a plateau was reached after 180 min and profiles were very similar up to 420 min, with most fragments ranging from 20 to 40 nt in length (Fig. 3B). An initial experiment was performed with time-points from 15 to 180 min, using 500ng of RNA for each point, followed by a separate follow-up experiment (using 155ng of RNA for each point and therefore shorter peaks) covering time-points from 180 to 420 min (Fig. 3B).

The transcriptional landscape of EVs revealed by the Total Transcriptome One-Step Protocol. rRNA was the most abundant class of RNAs in the EVs, corresponding to an average of 73% of total sequences in all samples. After filtering, an average of 1.2 million reads remained for each sample (Table 2). To account for the RNA fragmentation protocol and the miRNA size range, sequences with a length between 12–25 nt (an average of 332,549 reads/sample) were first mapped against miRBase v20 by using miRDeep2 software for miRNA

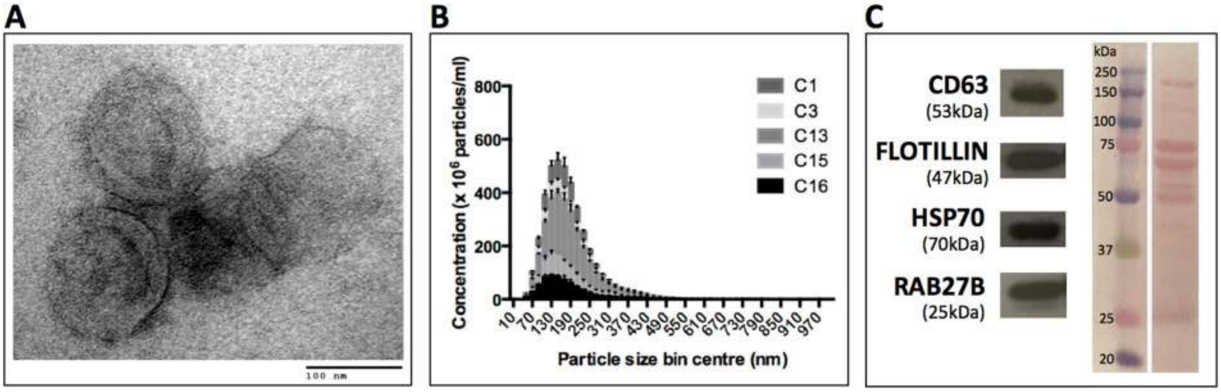

Figure 2. Characterization of EVs isolated from plasma by ultracentrifugation. (**A**) Whole-mount transmission electron micrograph of EVs displaying the characteristic "cup-shaped" morphology. (**B**) NanoSight quantification plot of EVs concentration in function of size shows that the majority of EVs are less than 250 nm in the healthy control samples (n = 5). (**C**) Western blot of the vesicle-associated markers CD63, FLOTILLIN, HSP70 and RAB27B from a pool of plasma samples (n = 4) and corresponding Ponceau staining of membrane before blocking.

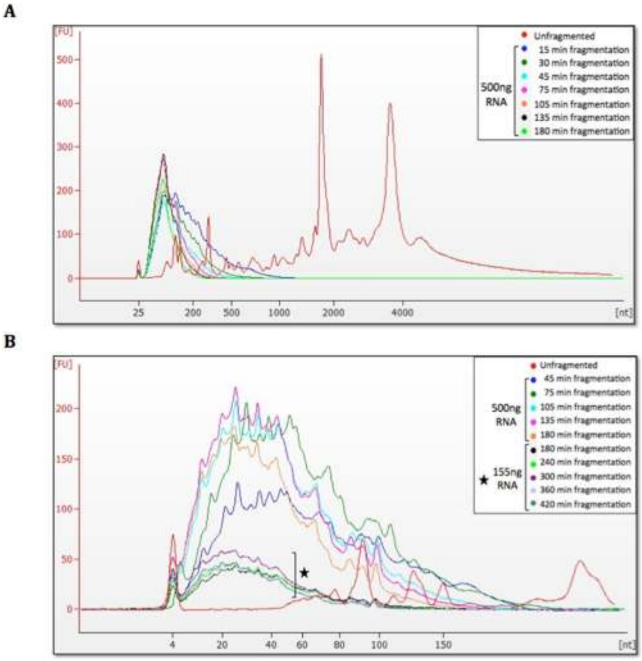

Figure 3. Bioanalyzer electropherogram analysis of fragmentation time-points of leukocyte RNA. Arbitrary fluorescence units (FU) are plotted as a function of RNA size in nucleotides (nt). (**A**) Analysis of fragmentation time-points 15–180 min with Agilent RNA 6000 Pico Kit shows the expected ribosomal RNA peaks in non-fragmented sample (in red), whereas after all fragmentation time-points the majority of RNAs are below 200 nt in size. (**B**) Analysis of fragmentation time-points from 45 to 420 min with Agilent Small RNA Kit shows that the majority of RNAs are smaller than 40 nt in size for all time-points, and that a plateau is reached after approximately 180 min, as no further reduction in size is observed with longer fragmentation times. For the longer fragmentation periods, reduced amounts of RNA (marked with a star) were used to better simulate the enzymatic kinetics in the presence of less (but still detectable) RNA, and the plateau region is still the same.

annotation. Reads not mapped by miRDeep2 and reads with lengths above 25 nt were then mapped to the hg19 reference genome by using STAR[22]. We obtained the coordinates of tRNAs from GtRNAdb (gtrnadb.ucsc.edu), and subsequently used BEDTools[27] to compare the genomic coordinates to find reads overlapping with tRNA. After annotating miRNAs and tRNAs, remaining reads representing other RNA classes contained in the Ensembl database were annotated according to biotypes defined in the Ensembl glossary (http://grch37.ensembl.org/info/website/glossary.html) as depicted in Table 3 and Fig. 1. Table 3 lists the number of different transcripts identified in each RNA biotype, considering only those transcripts identified by at least two reads. Protein coding was the most diverse category in EVs, followed by short noncoding RNAs.

1	2	3	4	5	6	7	8	9
Sample	Raw reads	Reads after FastQ screen filter	% of filtered reads	% of rRNA filtered reads	Reads 12–25 nt used for miRNA analysis (%)	miRBase mapped reads (%)	Reads used for other RNA analysis (%)	Ensembl mapped reads (%)
C1	18,085,235	1,653,832	90.9	75.9	458,186 (27.7)	33,157 (7.2)	1,620,675 (98.0)	576,866 (35.6)
C3	7,843,315	727,633	90.7	73.4	221,985 (30.5)	23,594 (10.6)	704,039 (96.8)	170,592 (24.2)
C13	13,830,831	1,148,352	91.7	74.8	291,302 (25.4)	14,734 (5.1)	1,133,618 (98.7)	358,268 (31.6)
C15	10,111,670	1,023,795	89.9	73.8	300,569 (29.4)	25,465 (8.5)	998,330 (97.5)	232,061 (23.2)
C16	10,936,417	1,509,424	86.2	69.3	390,705 (25.9)	40,801 (10.4)	1,468,623 (97.3)	526,181(35.8)
Average	12,161,494	1,212,607	89.9	73.4	332,549 (27.8)	27,550 (8.4)	1,185,057 (97.7)	372,793 (30.1)

Table 2. Summary of sequencing analysis of the total transcriptome of EVs isolated from the plasma of healthy controls. The % in columns 4 and 5 are relative to column 2; the % in columns 6 and 8 are relative to column 3; the % in column 7 is relative to column 6, and the % in column 9 is relative to column 8.

Sample	Protein coding	Pseudogene	Long noncoding	Short noncoding			
				miRNA mature (miRBase)	tRNA (GtRNAdb)	misc_RNA	other*
C1	7,824	91	220	765			
				181	220	218	146
C3	5,461	34	101	554			
				163	172	110	109
C13	7,937	83	207	738			
				157	259	189	133
C15	6,134	44	131	626			
				162	217	133	114
C16	8,660	171	211	887			
				216	235	254	182
Average	7,203	85	174	714			
				176	221	181	137

Table 3. Summary of the total number of distinct transcripts identified for different RNA categories in EVs isolated from the plasma of healthy controls. *The category 'other' includes: miRNA precursors (miRBase), miRNA (Ensembl), Mt_rRNA, Mt_tRNA, rRNA, snoRNA, snRNA.

Considering relative abundances based on the number of reads mapping to each RNA category, short noncoding RNA was the most represented biotype, corresponding to 73.25% of the mapped reads. The second most represented biotype was protein-coding RNAs, represented by 24.46% of the total mapped reads, followed by long noncoding RNAs with 2.16% and pseudogenes with 0.10% of reads (Fig. 4A). In the short noncoding RNA category, tRNAs were predominant, with 57.29% of mapped reads considering short RNAs identified by at least two reads, followed by mitochondrial rRNAs (Mt_rRNA) with 14.83%, miscellaneous RNAs (misc_RNA) with 13.23%, and miRNAs with 12.86% of mapped reads (Fig. 4B). Among the misc_RNAs, Y_RNAs and SRP_7SL_RNAs were predominant, with 57.45% and 39.44% of mapped reads, followed by Vault_RNAs and 7SK_RNAs with 2.99% and 0.11% of mapped reads, respectively (Fig. 4C). Miscellaneous RNAs have been previously described as enriched in EVs[11,14], and much remains to be understood about their regulatory functions. Similar distributions of RNA biotypes are also observed when analyses were performed with the more abundant transcripts represented by at least 10 reads (Fig. 4D–F), suggesting that our libraries, even with modest coverage after the rRNAs were filtered out, equally represented the distinct transcript classes.

For the analysis of reads mapping against longer transcripts (not miRNAs annotated by miRBase or tRNAs annotated by GtRNAdb), we implemented an additional filter to reduce potentially unreliable mapping due to short read length, which reflects mapped transcripts with poor and heterogeneous horizontal coverage. We used the Gini normalization coefficient, a measure of statistical dispersion often used to describe inequalities between parts, such as income distributions in a population[23]. In this analysis, all transcripts were divided in 10 equal parts independent of transcript size, and the Gini normalization coefficient was computed based on the total amount of reads aligning in each part. A Gini coefficient close to one indicates high inequality among the parts, which in our scenario translates to heterogeneous transcript coverage, suggesting alignment issues. After manual curation, considering the sequencing coverage obtained here, we elected a maximum Gini coefficient of 0.6 as a cut-off value based on the correlation of adequate horizontal coverage.

In Suppl. Table 2 we list the average expression levels for transcripts annotated in the Ensembl database, after excluding miRNAs annotated by miRBase and tRNAs annotated by GtRNAdb, and that passed the Gini coefficient filtering step, normalized to 100,000 reads/sample relative to the total counts of all mapped Ensembl transcripts. It can be observed that most of these (80.1%) were present in all five individuals and should, therefore, consist in the core EV-transcriptome.

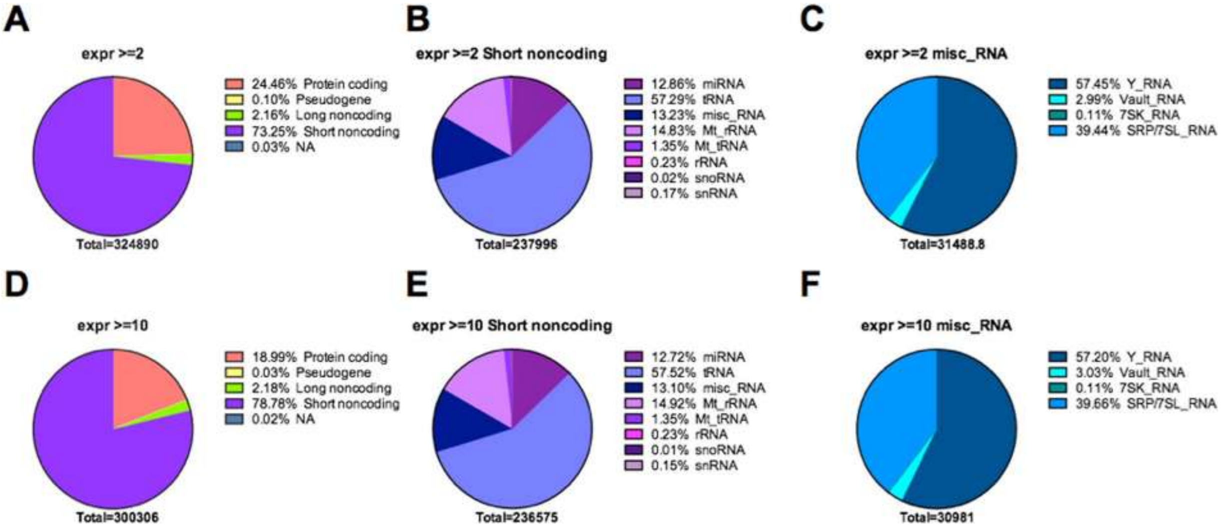

Figure 4. Distribution of the RNA classes identified in the sequencing analysis of the total transcriptome of EVs isolated from the plasma of healthy controls (average values for five samples). Pie charts show the distribution of mapped reads according to gene biotype, as defined by Ensembl, and sequencing coverage. Distribution of reads is shown for the four major Ensembl biotypes (**A**,**D**); the short noncoding biotype (**B**,**E**) and the misc_RNA biotype (**C**,**F**). Genes considered were covered by at least 2 reads (**A**–**C**), or by a minimum of 10 reads (**D**–**F**).

Notably, in addition to the RNA classes described above, our sequencing approach also enables the identification and quantification of backspliced circular RNAs (circRNAs), which have been recently identified in EVs[28,29]. After considering only the predicted RNA-circularization events that were confirmed by at least two reads, we found a total of 83 different putative circRNAs in EVs isolated from the plasma of healthy women, including 68/83 (82%) that are also present with the same genomic coordinates (start and end) in circBase[30–34] and 15 novel circRNA structures (Suppl. Table 3). Representative circRNAs were validated by PCR (by using outward primers to cover the circularizing point), followed by Sanger DNA sequencing (Suppl. Table 3, see an example in Suppl. Figure 1).

Confirmation of differential miRNA abundance and intravesicular location of transcripts. In order to verify the quantitative capabilities of our NGS-based protocol, we next selected one miRNA with high read counts (hsa-miR-223-3p - 313 reads/million) and one with low read counts (hsa-let-7g-5p - 19 reads/million). The expression value in reads/million was calculated by taking the number of reads mapped for each miRNA, multiplying by 1,000,000 and dividing by the total number of reads for that particular sample. This normalization coefficient was calculated for each sample and an average of the five samples is reported above. Quantitative RT-PCR validation of these miRNAs showed a ~12-fold difference in quantity calculated by 2^(delta Ct), comparable to a differential expression of ~16-fold calculated by NGS. At the same time, we conducted an experiment to confirm the intravesicular location of these same molecules by treating the EVs with RNase-A before RNA extraction. Three conditions were compared in parallel, starting with the same plasma volumes (0.5 ml) for each sample: i) mock-treatment of intact EVs, ii) intact EVs treated with RNase-A, and iii) disrupted EVs treated with RNase-A. Whereas the Ct values for intact EVs – mock- and RNase A-treated – were quite similar (target reduction varying from 1.32-fold to 3.13-fold after RNAse treatment), the treatment of lysed EVs showed a drastic Ct increment, corresponding to a 64.3-525.8-fold reduction in miRNA quantities when compared to intact EVs, strongly suggesting that most RNAs represented in our protocol are indeed intra-vesicular (Fig. 5).

Discussion

Extracellular vesicle analysis is an essential component of liquid biopsies in contemporary research and medical applications. Thus, the dissection of EV-cargo from human samples and animal models has the potential to reveal elements that may be key for tissue homeostasis and tissue-tissue communication. In cancer patients, the information provided by plasma-derived EVs may contain valuable biomarkers for precision medicine, complementing the value of other biomarkers found in circulating tumor cells or circulating tumor DNA, including molecules that can play a role in priming tissues that will receive metastatic cells[35]. In this sense, the capability of defining the total EV transcriptome is remarkably relevant.

Here we present a new approach: the One-Step Total Transcriptome Protocol, which, for the first time, allowed the description of all known biotypes of plasma derived EV-RNAs in a single NGS library. A major technical hurdle found during NGS analysis of limited RNA input samples is the formation of an excess of adaptor-dimers during library preparation, a technical issue that is present when the available amount of RNA is <200 ng[26,36]. In this sense, our protocol of RNA fragmentation by RNase-III treatment allowed the increment of the number of molecules available for ligation to adaptors, with no need to amplify the initial material. This feature counteracted the artifactual formation of adaptor-dimers and avoided possible amplification and NGS-representation biases of the heterogeneous vesicular transcriptome.

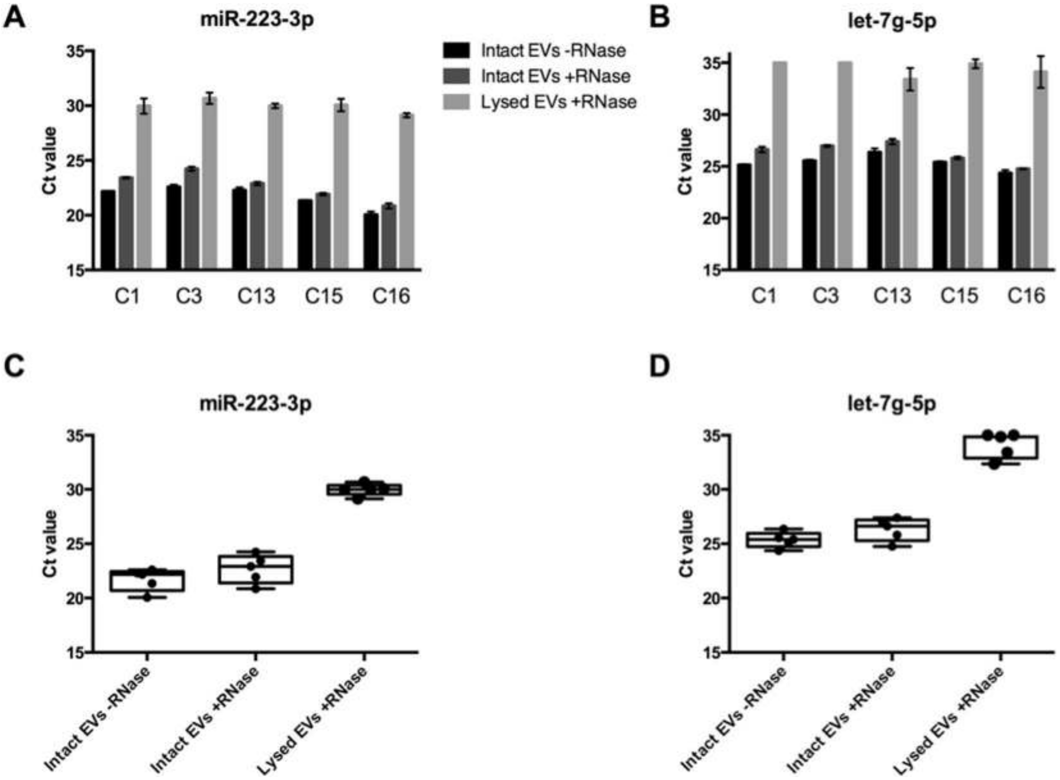

Figure 5. Quantitative RT-PCR analysis of two miRNAs (miR-223-3p and let-7g-5p) identified by sequencing. The treatment of intact EVs with RNase before RNA extraction minimally altered Ct values of both miRNAs, strongly suggesting that these miRNAs were derived from the EVs-cargo and thereby protected from digestion by the membrane bilayer, whereas there was a large increase in Ct values when lysed EVs were similarly treated. Undetermined values (no detection by qRT-PCR) were assigned as Ct = 35. (**A**,**B**) Bar-graphs for miR-223-3p and let-7g-5p, showing Ct values for each treatment per sample; (**C**,**D**) Box-plots for miR-223-3p and let-7g-5p, showing Ct values per treatment (dots are results from each one of the five samples).

Previous results from our group (data not shown) also obtained from human plasma-derived EVs isolated with this same ultracentrifugation protocol, but using no fragmentation step, resulted in 46–82% of the reads to be adaptor-dimers, having as input the total RNA isolated from 5 ml of plasma-derived EVs. After the adoption of the protocol introduced here, the frequency of adaptor-dimers was reduced to 13–38%, by using RNA from 2.5 ml of plasma. The enzymatic digestion of the whole RNA-content of EVs, reducing all transcripts to a size range of small RNAs but still permitting their precise annotation and quantification, allowed us to optimize the library preparation protocol intended for small RNAs, and yet to represent other RNAs subtypes, and to ultimately cover essentially the whole EV-transcriptome with a single library.

This protocol revealed the presence of thousands of transcripts from distinct RNA classes inside the vesicles, including long RNAs and circRNAs. The fragmentation of these molecules, with an original average size of 2.5–3.5Kb, to an average size of 38 nt allowed us to increase the number of these molecules at least 65-fold, negatively impacting the formation of undesirable adaptor-dimers. Besides reducing the amounts of adaptor-dimers, the benefits of this approach include i) the reduction of reagents amount, time and sequencing costs to less than half, ii) the ability to use low sample volume input and iii) the unbiased analysis of the entire transcriptome at once, without the inherent prejudice of methodologies based on size selection or template pre-amplification. In this sense, it is well worth mentioning that a recent large-scale RNA-seq study showed a clear batch effect in the percentage distribution of RNA species identified due to variations in the gel size selection step – 40 gels were run during the library preparation of 192 samples and the relative quantities of miRNAs subsequently identified were consistently higher in the first 19 gels compared to the remaining 21 gels, where more piwiRNAs were identified[12].

Another limitation that may derive from the fragmentation of the molecules is the sometimes-ambiguous mapping of some reads to mRNAs. However, this matter was solved with a more stringent bioinformatics pipeline for mapping short reads, incorporating the Gini normalization coefficient as a means to indicate transcripts with good and homogeneous horizontal coverage as an extra step to essentially eliminate ambiguous mapping. Moreover, the Gini coefficient would be important to be implemented in further analysis where the evaluation of differential expression among groups of samples is a goal of the study.

For each sample of the study we generated an average of 12.1 million reads (Table 2). After filtering with FastQ screen, about 73% of the EVs transcriptome showed to be derived from human rRNA genes. The field has been somewhat controversial regarding the presence or absence of rRNA in EVs, with several studies claiming their absence due to the lack of the characteristic 18S and 28S rRNA peaks in Bioanalyzer traces[3,7,37,38]. However, these peaks only reflect the presence of intact rRNA molecules and recent whole transcriptome RNA-seq studies (similar to ours) have found up to 97% of rRNA and indeed showed that EVs are enriched in rRNAs fragments, which are not visible in Bioanalyzer profiles[16]. One might emphasize that for clinical patient-derived samples, as of today, the use of rRNA depletion kits is usually not feasible due to the requirement of high amounts of starting RNA mass (>100 ng), which are far beyond the RNA-content available in typical clinical samples (<1 ng). RNA-seq studies that performed size selection procedures (either gel- or bead-based) might perhaps be able to circumvent the rRNAs detection by enriching for a smaller RNA fraction. However, this would likely lead to a biased view of the EVs-cargo, as a substantial population of mRNAs and long noncoding RNAs is bound to be also excluded together with the rRNA sequences, due to an overlap in transcript length.

The transcriptional landscape of circulating EVs in these subjects revealed that the majority of sequences, besides the rRNAs, correspond to short noncoding RNAs. In this category, the major players are tRNAs, miscRNAs and miRNAs. Among the miscRNAs found in EVs, we highlight the abundance of Y-RNAs, noncoding molecules that associate with replicating euchromatin[39], have been associated with breast cancer[40], atherosclerosis-related diseases[41] and are enriched in EVs[11,14] (Fig. 4). Among the top-50 most abundant transcripts identified in EVs (Suppl. Table 2), 38% are not described in EV-cargo databases[42,43], which do not cover lncRNAs. Notably, Vesiclepedia[42] has only seven studies with RNA analysis of plasma/serum samples, and none of these employed NGS technology for EVs characterization. Also, only two out of the seven studies analyzed mRNA profiles, all other studies focused solely on miRNAs.

Our analysis also confirmed the presence of back-spliced circRNAs in EVs. These transcripts - characterized by the presence of some exons spliced in non-canonical order due to the circularization of an otherwise linear transcript - have been found to be stable, abundant, evolutionarily conserved and in some cases these molecules appear to act as miRNA sponges that can block miRNA-regulation of their linear counterpart-genes in recipient cells[28–30,44–46]. Our findings based in the consensus alignment of high-quality reads indicated the presence of 83 circRNAs in these individuals, including 15 new circRNAs, of which seven have the same start or end positions of previously described circRNAs present in circBase, suggesting possible new isoforms and eight that were not previously annotated (Suppl. Table 3).

Our qRT-PCR validation of the NGS results demonstrated the intra-vesicular location of the isolated RNA and suggested that their quantification by NGS, given by read-counts derived from the protocol presented here, does not introduce significant quantification biases, which is relevant for the study of potential biomarkers in pathological conditions. The protocol presented here enables for the first time the analysis of total transcriptomes when only limiting amounts of RNA are available, as in the case of extracellular vesicles, allowing the representation of all RNA classes while also preserving the quantitative differences between molecules. As such, we conclude that this One-Step methodology might become a standard in the emerging field of liquid biopsy applications.

References

1. Barteneva, N. S. *et al*. Circulating microparticles: square the circle. *BMC Cell Biol* **14**, 23 (2013).
2. Colombo, M., Raposo, G. & Théry, C. Biogenesis, Secretion, and Intercellular Interactions of Exosomes and Other Extracellular Vesicles. *Annu Rev Cell Dev Biol* **30**, 255–289 (2014).
3. Valadi, H. *et al*. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat Cell Biol* **9**, 654–9 (2007).
4. Amorim, M. *et al*. The overexpression of a single oncogene (ERBB2/HER2) alters the proteomic landscape of extracellular vesicles. *Proteomics* **14**, 1472–9 (2014).
5. Abels, E. R. & Breakefield, X. O. Introduction to Extracellular Vesicles: Biogenesis, RNA Cargo Selection, Content, Release, and Uptake. *Cell Mol Neurobiol* **36**, 301–312 (2016).
6. Al-Nedawi, K. *et al*. Intercellular transfer of the oncogenic receptor EGFRvIII by microvesicles derived from tumour cells. *Nat Cell Biol* **10**, 619–24 (2008).
7. Skog, J. *et al*. Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nat Cell Biol* **10**, 1470–6 (2008).
8. Deregibus, M. C. *et al*. Endothelial progenitor cell derived microvesicles activate an angiogenic program in endothelial cells by a horizontal transfer of mRNA. *Blood* **110**, 2440–8 (2007).
9. Ahsan, N. A. *et al*. Presence of Viral RNA and Proteins in Exosomes from Cellular Clones Resistant to Rift Valley Fever Virus Infection. *Front Microbiol* **7**, 139 (2016).
10. Huang, X. *et al*. Characterization of human plasma-derived exosomal RNAs by deep sequencing. *BMC Genomics* **14**, 319 (2013).
11. Nolte'T Hoen, E. N. M. *et al*. Deep sequencing of RNA from immune cell-derived vesicles uncovers the selective incorporation of small non-coding RNA biotypes with potential regulatory functions. *Nucleic Acids Res* **40**, 9272–9285 (2012).
12. Yuan, T. *et al*. Plasma extracellular RNA profiles in healthy and cancer patients. *Sci Rep* **6**, 19413 (2016).
13. Quek, C. *et al*. Defining the purity of exosomes required for diagnostic profiling of small RNA suitable for biomarker discovery. *RNA Biol* 00–00 https://doi.org/10.1080/15476286.2016.1270005 (2016).
14. Lefebvre, F. A. *et al*. Comparative transcriptomic analysis of human and Drosophila extracellular vesicles. *Sci Rep* **6**, 27680 (2016).
15. San Lucas, F. A. *et al*. Minimally invasive genomic and transcriptomic profiling of visceral cancers by next-generation sequencing of circulating exosomes. *Ann Oncol* **27**, 635–641 (2016).
16. Jenjaroenpun, P. *et al*. Characterization of RNA in exosomes secreted by human breast cancer cell lines using next-generation sequencing. *PeerJ* **1**, e201 (2013).
17. Schageman, J. *et al*. The complete exosome workflow solution: From isolation to characterization of RNA cargo. *Biomed Res Int* **2013**, (2013).
18. Li, M. *et al*. Analysis of the RNA content of the exosomes derived from blood serum and urine and its potential as biomarkers. *Philos Trans R Soc B Biol Sci* **369**, 20130502–20130502 (2014).
19. Théry, C. *et al*. Isolation and characterization of exosomes from cell culture supernatants and biological fluids. *Curr Protoc cell Biol* Chapter 3, 3.22.1-3.22.29 (2006).
20. Andrews, S. No Title http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/ (2011). Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/.
21. Mackowiak, S. D. Identification of novel and known miRNAs in deep-sequencing data with miRDeep2. *Curr Protoc Bioinforma* Chapter 12, Unit12.10. (2011).
22. Dobin, A. *et al*. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

23. Gini, C. Measurement of Inequality of Incomes. *Econ J* **31**, 124 (1921).
24. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20**, 1666–70 (2014).
25. Sanguinetti, C. J., Dias Neto, E. & Simpson, A. J. Rapid silver staining and recovery of PCR products separated on polyacrylamide gels. *Biotechniques* **17**, 914–21 (1994).
26. Head, S. R. *et al*. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* **56**, 61–4, 66, 68, passim (2014).
27. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
28. Li, Y. *et al*. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res* **25**, 981–984 (2015).
29. Lasda, E. & Parker, R. Circular RNAs co-precipitate with extracellular vesicles: A possible mechanism for circrna clearance. *PLoS One* **11**, 1–11 (2016).
30. Jeck, W. R. *et al*. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–57 (2013).
31. Memczak, S. *et al*. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–8 (2013).
32. Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. & Brown, P. O. Cell-type specific features of circular RNA expression. *PLoS Genet* **9**, e1003777 (2013).
33. Zhang, Y. *et al*. Circular intronic long noncoding RNAs. *Mol Cell* **51**, 792–806 (2013).
34. Rybak-Wolf, A. *et al*. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol Cell* **58**, 870–85 (2015).
35. Peinado, H. *et al*. Pre-metastatic niches: organ-specific homes for metastases. *Nat Rev Cancer* **17**, 302–317 (2017).
36. Buschmann, D. *et al*. Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow. *Nucleic Acids Res* **44**, 5995–6018 (2016).
37. Cheng, L., Sharples, R. A., Scicluna, B. J. & Hill, A. F. Exosomes provide a protective and enriched source of miRNA for biomarker profiling compared to intracellular and cell-free blood. *J Extracell vesicles* **3**, 1–14 (2014).
38. Ji, H. *et al*. Deep Sequencing of RNA from Three Different Extracellular Vesicle (EV) Subtypes Released from the Human LIM1863 Colon Cancer Cell Line Uncovers Distinct Mirna-Enrichment Signatures. *PLoS One* **9**, e110314 (2014).
39. Kheir, E. & Krude, T. Non-coding Y RNAs associate with early replicating euchromatin in concordance with the origin recognition complex. *J Cell Sci* **130**, 1239–1250 (2017).
40. Dhahbi, J. M., Spindler, S. R., Atamna, H., Boffelli, D. & Martin, D. I. Deep Sequencing of Serum Small RNAs Identifies Patterns of 5′ tRNA Half and YRNA Fragment Expression Associated with Breast Cancer. *Biomark Cancer* **6**, 37–47 (2014).
41. Repetto, E. *et al*. RNY-derived small RNAs as a signature of coronary artery disease. *BMC Med* **13**, 259 (2015).
42. Kalra, H. *et al*. Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation. *PLoS Biol* **10**, e1001450 (2012).
43. Kim, D.-K. *et al*. EVpedia: an integrated database of high-throughput data for systemic analyses of extracellular vesicles. *J Extracell vesicles* **2**, 20384 (2013).
44. Hansen, T. B. *et al*. Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–8 (2013).
45. Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. & Brown, P. O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* **7**, e30733 (2012).
46. Piwecka, M. *et al*. Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. *Science* https://doi.org/10.1126/science.aam8526 (2017).

Acknowledgements

This work received financial support from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP grants 2011/09172-3 & 2014/26897-0). EDN is a fellow from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and from Associação Beneficente Alzira Denise Hertzog Silva (ABADHS). RP and WA acknowledge the Gillson-Longenbaugh Foundation for support. Work in Dr. Calin's laboratory is supported by National Institutes of Health (NIH/NCATS) grant UH3TR00943-01 through the NIH Common Fund, Office of Strategic Coordination (OSC). We kindly thank Dr. Thiago C. Chulam for facilitating the recruitment of subjects during the cancer prevention campaign at ACCCC and nurse Fernanda Sulian de Carvalho who helped with blood sample collection, as well as the technicians André Haraguti Aguillera (UNIFESP) and Roberto Cabado Modia Jr. (USP) for their help with electron microscopy. We thank Dr. Helen Pickersgill (Life Science Editors) for professional editorial assistance.

Author Contributions

M.G.A., E.D.N. and D.N.N. designed the experiments and wrote the manuscript; M.G.A. performed the experiments; R.V., R.D. and I.T.S. performed bioinformatic analyses of N.G.S. data; M.P.P. helped with subject recruitment and sample collection; V.F. and R.S.C. helped with electron microscopy analysis; M.G.A., G.A.C., R.P., W.A., I.T.S., E.D.N. and D.N.N.: data review and manuscript editing. All authors have reviewed and accepted the final manuscript version.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-14264-5.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.