Unbiased RNA-Seq-driven identification and validation of reference genes for quantitative RT-PCR analyses of pooled cancer exosomes

Yao Dai
Tongji University
Yumeng Cao
Tongji University
Jens Köhler
Dana Farber Cancer Institute Department of Medical Oncology
Aiping Lu
Tongji University
Shaohua Xu
Tongji University
Haiyun Wang (wanghaiyun@tongji.edu.cn)
Tongji University

Research article

Keywords: reference gene, qRT-PCR, cancer exosome, RNA-Seq, miRNA-Seq

Posted Date: December 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-29893/v3

Version of Record: A version of this preprint was published on January 6th, 2021. See the published version at https://doi.org/10.1186/s12864-020-07318-y.

Unbiased RNA-Seq-driven identification and validation of reference genes for quantitative RT-PCR analyses of pooled cancer exosomes

Yao Dai[1], Yumeng Cao[1], Jens Köhler[2], Aiping Lu[1], Shaohua Xu[1,*], Haiyun Wang[1, *]

[1]*Department of Gynecology, Shanghai First Maternity and Infant Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China*

[2]*Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA*

[*] Corresponding author

E-mail: daiyao@tongji.edu.cn (Dai Y),

yumengcao163@163.com (Cao Y),

jens_kohler@dfci.harvard.edu (Köhler J)

APLV@tongji.edu.cn (Lu A),

xushaohua@tongji.edu.cn (Xu S),

wanghaiyun@tongji.edu.cn (Wang H)

Abstract

Background: Exosomes are extracellular vesicles (EVs) derived from endocytic compartments of eukaryotic cells which contain various biomolecules like mRNAs or miRNAs. Exosomes influence the biologic behaviour and progression of malignancies and are promising candidates as non-invasive diagnostic biomarkers or as targets for therapeutic interventions. Usually, quantitative real-time polymerase chain reaction (qRT-PCR) is used to assess gene expression in cancer exosomes, however, the ideal reference genes for normalization yet remain to be identified.

Results: In this study, we performed an unbiased analysis of high-throughput mRNA and miRNA-sequencing data from exosomes of patients with various cancer types and identify candidate reference genes and miRNAs in cancer exosomes. The expression stability of these candidate reference genes was evaluated by the coefficient of variation "CV" and the average expression stability value "M". We subsequently validated these candidate reference genes in exosomes from an independent cohort of ovarian cancer patients and healthy control individuals by qRT-PCR.

Conclusions: Our study identifies *OAZ1* and *hsa-miR-6835-3p* as the most reliable individual reference genes for mRNA and miRNA quantification, respectively. For superior accuracy, we recommend the use of a combination of reference genes - *OAZ1*/*SERF2*/*MPP1* for mRNA and *hsa-miR-6835-3p*/*hsa-miR-4468-3p* for miRNA analyses.

KEYWORDS: reference gene; qRT-PCR; cancer exosome; RNA-Seq; miRNA-Seq

Background

Exosomes are a class of extracellular vesicles (EVs) which are secreted by eukaryotic cells. Exosomes contain biomolecules, such as DNA, RNA, miRNA or proteins and are considered important mediators of intercellular communication [1-7]. Cancer cell-derived exosomes play a pivotal role in tumorigenesis and cancer progression as they modulate cancer cell biology, the tumor microenvironment and the immune response [7-13]. Tumor-derived exosomes can also be harnessed as non-invasive diagnostic biomarkers due to their abundance in biological fluids and the enrichment of tumor-relevant biomolecules such as mRNAs or miRNAs within [4, 14-17]. In the past, various exosome-based liquid biopsies studies have suggested clinical feasibility for cancer diagnosis [18-20].

To accurately explore exosomes as non-invasive biomarkers and to better understand their impact on cancer progression, the precise quantification of biomolecule abundance within exosomes is of utmost importance. Quantitative real-time polymerase chain reaction (qRT-PCR) is the most widely used laboratory technique to evaluate cell-intrinsic and exosomal gene expression patterns [21]. qRT-PCR offers the advantage of high sensitivity and specificity combined with reproducibility and low template input requirements [22, 23]. However, technical or experimental factors inherent to qRT-PCR, such as variable template integrity or efficiency of reverse transcription, can reduce the diagnostic accuracy [23-26]. In addition, the numbers, sizes, and compositions of exosomes are usually affected by many factors including the methodologies for exosome isolation, intracellular biological processes, cell culture parameters and the treatments of the parental cells, which introduce the difficulty for the characterization of the composition in exosomes [27-29]. To account

for this, reference genes with stable expression across different conditions or cancer subtypes are used to normalize gene expression [22, 30, 31]. Currently, the reference genes used for expression analyses in exosomes are most frequently those which are also used for tissue or cell lines, such as *ACTB*, *18S rRNA* and *GAPDH* [5, 32, 33]. Notwithstanding their broad use, expression levels of these housekeeping genes are not universally stable, thus decreasing the quantitative accuracy in exosome studies [22, 31, 34-37]. For example, the small nucleolar RNA *RNU6* is frequently used as a reference gene for miRNA expression analyses within cells [38-40], but the molecule is only expressed in the cell nucleus and not detected in exosomes [41-43]. Whereas some studies reported *RNU6* to be detectable in exosomes, this is most likely due to contamination of the exosome fraction with intact cells or cell debris [44, 45]. Therefore, the unvalidated use of classical housekeeping genes suitable for cell lines or tissues needs to be critically considered for the analysis of exosomes.

To address this unmet need of an unbiased identification and validation of reference genes or miRNAs for exosome studies, here, we performed a sequencing-driven analysis with high-throughput mRNA- and miRNA-Seq datasets from serum exosomes of patients with frequent cancer types and of healthy control individuals and subsequently validate these candidates by qRT-PCR in serum exosomes of an independent cohort of ovarian cancer patients and of healthy control individuals.

Results

Identification of candidate reference genes by an unbiased integrative analysis of pooled cancer mRNA-Seq datasets

To identify reference genes with stable expression in serum exosomes, we interrogated RNA-Seq data from 47 serum exosome samples of patients with PAAD, CRC and HCC as well as of 32 healthy control individuals (HC) and applied Deseq2 to evaluate expression levels across samples. Only genes with high expression in both, serum exosomes of cancer patients and of healthy individuals (measured as transcripts per million (TPM)) compared to the average gene expression level (pooled-transcriptome) were considered as potential reference candidates. Our analysis firstly identified 112, 117, and 85 stably expressed genes respectively in serum exosomes of PAAD, CRC and HCC (p value > 0.1), by comparing their patients with healthy control individuals using Deseq2 analysis. Then 48 genes were found to be universally stably expressed in serum exosomes of all cancers. By sorting these genes by their expression level, we identified ten highly expressed candidate reference genes (ADP-ribosylation factor 1 (*ARF1*), beta-2-microglobulin (*B2M)*, H3 histone pseudogene 6 (*H3F3AP4*), integral membrane protein 2B (*ITM2B*), membrane palmitoylated protein 1 (*MPP1*), ornithine decarboxylase antizyme 1 (*OAZ1*), protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 1 (*PCMTD1*), superoxide dismutase 2 (*SOD2*), small EDRK-rich factor 2 (*SERF2*), and WAS/WASL Interacting Protein Family Member 1 (*WIPF1*) (Fig. 1A, indicated by red dots and Tab. 1). The diagonal scatterplot distribution of candidate reference genes indicates consistent expression abundance between exosomes of cancer patients and of healthy control individuals (Fig. 1A), with a correlation coefficient of R=0.995. Furthermore, expression

patterns of candidate reference genes identified by the pooled cancer analysis (including PAAD, CRC and HCC) were recapitulated in each cancer subtype as well (Fig. 1B-D).

Evaluation of expression levels and stability of candidate reference genes

To further validate our predicted candidate reference genes for exosomes, we compared their respective expression levels and stabilities with those of nine classical housekeeping genes: beta-actin (*ACTB*), beta-2-microglobulin (*B2M*), ribosomal protein L13A (*RPL13A*), tyrosine 3-monooxygenase/tryptophane 5-monooxygenase activation protein zeta (*YWHAZ*), glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), vimentin (*VIM*), peptidylprolyl isomerase A (*PPIA*), aldolase A (*ALDOA*), and ubiquitin C (*UBC*). Overall, abundance of exosomal candidate reference genes (Fig. 2A) was similar to those of classical housekeeping genes (Fig. 2B). *B2M* had by far the highest overall expression abundance of all candidate reference genes (Fig. 2A) which was only surpassed by the classical housekeeping gene *ACTB* (Fig. 2B). We then assessed the expression stability across samples and tumor types by two measures: 1) the coefficient of variation "CV" as the standard deviation divided by the mean of the expression levels (transcripts per million - TPM), and 2) the average expression stability "M" determined by the geNorm algorithm. "CV" values for the exosomal candidate reference genes (0.405 to 0.723) (Fig. 2C) were significantly lower than those for classical housekeeping genes ($p=8.10e-04$, Wilcoxon rank-sum test) (Fig. 2D) with "M" values below 1.0, thus indicating higher expression stability across samples and tumor types (Fig. 2E). The "M" values were also significantly lower in candidate reference genes compared to those for classical housekeeping genes ($p=0.0279$, Wilcoxon rank-sum test) (Fig. 2F). The candidate reference genes were then sorted according to their expression stability from highest to lowest, and both, the "CV" and "M" criteria achieved similar ranks for most candidates. *OAZ1* was identified as the

gene with the highest expression stability across samples and tumor types (Tab. 1). We also identified and validated ten candidate reference genes respectively for each cancer subtype including PAAD (*FTL*, *OAZ1*, *FYB1*, *SERF2*, *SOD2*, *PCMTD1*, *ARPC2*, *NCOA4*, *HCLS1* and *TYROBP*), CRC (*B2M*, *RPL41*, *SNCA*, *RPS9*, *BTF3*, *ADIPOR1*, *HEMGN*, *SOD2*, *PCMTD1* and *NCOA4*), and HCC (*FTL*, *OAZ1*, *CD74*, *DDX5*, *PCMTD1*, *HCLS1*, *LSP1*, *RPL9*, *WIPF1* and *H3F3AP4*) as well (Suppl. Fig. 1).

Validation of candidate reference genes in exosomes of an independent cohort of ovarian cancer patients

Based on the promising results from the pooled analysis of serum exosomes of patients with different tumour types, we expected our predicted candidate reference genes to be applicable to serum exosomes from patients with various other cancer types as well. Therefore, we next sought to validate the candidate reference genes in a "real-life setting" in samples of serum exosomes of ten ovarian cancer patients and of ten healthy control individuals. The qRT-PCR results showed that as expected from the RNA-Seq data, *B2M* had the highest expression abundance among all candidates (Fig. 3A). Moreover, absolute abundance of *SOD2*, *H3F3AP4*, *OAZ1*, and *SERF2* were comparable to the expression level of *18S rRNA*, whereas the abundance of the remaining five genes (*ITM2B*, *ARF1*, *PCMTD1*, *WIPF1*, *MPP1*) was lower (Fig. 3A). Interestingly, the abundance of the reference candidate genes in serum exosomes of healthy control individuals and of ovarian cancer patients were highly consistent (Fig. 3A). Most candidate genes also exhibited high expression stability in ovarian cancer and healthy control individuals with "M" and "CV" values lower than 1.0 (Fig. 3B-E), even though some variation occurred regarding the gene order between both stability indicators. Whereas *MPP1*, *WIPF1*, *SOD2* and *OAZ1* exhibited lower "CV" values in exosomes of healthy individuals (Fig. 3C), in both

exosome groups, *OAZ1* had the lowest "M" value (Fig. 3D-E). The "M" values for *OAZ1*, *ITM2B*, *SERF2*, *MPP1*, *H3F3AP4*, and *ARF1* were advantageous over *18S rRNA*, whereas *WIPF1*, *B2M*, *SOD2* and *PCMTD1* in part had clearly higher "M" values indicating reduced expression stability (Fig. 3D). The expression stability of *18S rRNA* was lower (indicated by a higher "M" value") compared to many of the identified candidate reference genes especially in exosomes of healthy control individuals (Fig. 3D-E).

To quantify gene expression levels more accurately, multiple reference genes can be used [46]. Therefore, we also determined the expression stability of respective combinations of candidate reference genes by determining the average gene-specific variation with the geNorm algorithm for RNA-Seq datasets in exosomes of the pooled cancer populations and for qRT-PCR data of exosomes from ovarian cancer patients. Overall, three combinations according to their expression stability ranks (Table 1) were evaluated: 1) genes 1-3 (*OAZ1*, *SERF2*, *MPP1*); 2) genes 4-6 (*H3F3AP4*, *WIPF1*, *PCMTD1*); and 3) genes 8-10 (*SOD2*, *B2M*, *ITM2B*). The first group with a combination of *OAZ1*, *SERF2* and *MPP1* had the lowest average gene-specific variations in exosomes of the pooled patient group including PAAD, HCC and CRC (RNA-Seq, Suppl. Fig. 2A) as well as in ovarian cancer patients (qRT-PCR, Suppl. Fig. 2B) indicating the highest expression stability.

Identification and validation of candidate reference miRNAs in cancer exosomes

In addition to mRNA, exosomes also contain miRNA. To identify reliable miRNAs for normalization in exosomes, we analyzed miRNA-Seq data of 72 serum exosome samples of patients with HCC, HNSCC, LCA, NBL, OVA, and THCA and 31 serum exosome samples of healthy control individuals. We identified six candidate reference miRNAs with

high and stable expression: hsa-*miR-125-5p, hsa-miR-192-3p*, *hsa-miR-4468*, *hsa-miR-4469, hsa-miR-6731-5p*, and *hsa-miR-6835-3p* (Fig. 4A). Expression levels and stability of the candidate reference miRNAs were evaluated in the exosomes of pooled cancer and further validated in the exosomes of ovarian cancer and healthy control individuals (Fig. 4B-J). Across the pooled exosomes of six cancer types, but also for each individual cancer type, these candidate miRNAs show high expression and similar abundance compared to exosomes of healthy control individuals (depicted as counts per million (CPM)) (Fig. 4B, Suppl. Fig. 3). Among all candidate miRNAs, *hsa-miR-6835-3p* had the highest expression level across samples and tumor types (Tab. 2). And *hsa-miR-4468* had the highest and *hsa-miR-6731-5p* the lowest expression stability across samples and cancer types as indicated by low and high "CV" and "M" values, respectively (Fig. 4E, H). Overall, "M" values for all candidate miRNAs were low (<1.5), indicating their general expression stability and potential utility as candidate reference miRNAs for exosomes. By integrating both stability indicators "CV" and "M", candidate reference miRNAs were ranked and hsa-miR-4468 showed the highest overall expression stability across samples and tumor types (Tab. 2). Finally, *hsa-miR-6835-3p with* high expression level and stability was identified as the best reference miRNA.

To further validate the predicted reference miRNA candidates, we measured their expression levels by qRT-PCR in serum exosomes of patients with ovarian cancer (n=10) and of healthy control individuals (n=10). miRNA abundance was calculated as cycle threshold numbers (Ct) relative to *ce-miR-39-1*. *ce-miR-39-1* is a frequently used miRNA for normalization (Fig. 4C-D). These results showed the highest expression for

206 *hsa-miR-4469* in exosomes of ovarian cancer patients even though all miRNAs were less
207 abundant than *ce-miR-39-1* (Fig. 4C-D). In exosomes of ovarian cancer patients,
208 *hsa-miR-4469* and *hsa-miR-4468* were the miRNAs with the highest and lowest expression
209 levels, reproducing the results for exosomes of healthy control individuals (Fig. 4C-D).
210 Compared to the miRNA-Seq analysis (Fig. E, H), *hsa-miR-6731-5p*, *hsa-miR-4468*,
211 *hsa-miR-192-3p* and *hsa-miR-6835-3p* exhibited lower "CV" and "M" values indicating
212 even higher expression stability in a "real-life" setting (Fig. 4F, G, I, J). Overall, all
213 candidate reference miRNAs in exosome of ovarian cancer and healthy control individuals
214 exhibited "M" values smaller than 1.5 indicating high expression stability (Fig. 4I-J).

215

216 Furthermore, the expression stability of combinations of multiple reference miRNAs
217 was determined by the average gene-specific variation. We generated three combinations
218 of two candidate reference miRNAs each according to their expression stability ranks (Tab.
219 2): 1) miRNAs 1-2 (*hsa-miR-4468* and *hsa-miR-6835-3p*), 2) miRNAs 3-4
220 (*hsa-miR-192-3p* and *hsa-miR-125a-5p)*, and 3) miRNAs 5-6 (*hsa-miR-4469* and
221 *hsa-miR-6731-5p*). The combination of *hsa-miR-6835-3p* and *hsa-miR-4468* had the
222 highest expression stability in exosomes of pooled groups of patients affected by PAAD,
223 HCC and CRC (miRNA-Seq data, Suppl. Fig. 4A) or by ovarian cancer (qRT-PCR data,
224 Suppl. Fig. 4B).

225

226 **Discussion**

Exosomes are nano-sized (<200nm in diameter) biovesicles which are released into the surrounding body fluids upon fusion of endocytic compartments with the plasma membrane[47] . Exosomes transfer various types of cargo from donor to acceptor cells among them nucleic acids, mRNAs and miRNAs were the first nucleic acids to be identified in exosomes [3]. Interestingly, some mRNAs and miRNAs are even specifically enriched in cancer exosomes implying a critical role for cancer biology and progression. Therefore, exosomes can be harnessed as diagnostic biomarkers or as targets for therapeutic interventions [3, 5, 48-50]. To characterize the composition of exosomes, the accurate quantification of mRNA and miRNA expression within the exosome fraction is critical. qRT-PCR combines high sensitivity and specificity with high reproducibility and low template input requirements and is therefore an ideal technology for exosome studies [22, 23]. qRT-PCR analyses, however, require the selection of appropriate reference genes to avoid variation in gene expression results under different experimental conditions (e.g. tumor cell vs. exosome) [22, 30, 31, 51] and currently, the ideal reference genes for the analysis of exosomes across cancers or for comparison of expression with cancer cells or tissues remain largely unknown [52, 53]. Often, classical housekeeping genes used for gene expression analyses in tissues or cell lines are used for exosome studies as well, but the expression stability of these genes is not unconditionally guaranteed for exosome samples thereby limiting the analytical accuracy. In this context, previous studies have confirmed that there is no universal reference gene for normalization under different conditions [35, 36, 54, 55].

Therefore, here, we sought to perform an unbiased and sequencing-driven analysis of publicly available high-throughput RNA- and miRNA-Seq datasets to identify and experimentally validate reference genes and combinations of these genes to generate accurate RNA- and miRNA-expression data from serum exosomes of cancer patients. From the pooled RNA- and miRNA-Seq datasets we identify multiple potential candidate reference genes in cancer exosomes (*ARF1*, *B2M*, *H3F3AP4*, *ITM2B*, *MPP1*, *OAZ1*, *PCMTD1*, *SOD2*, *SERF2*, *WIPF1* for mRNA analyses and *hsa-miR-125-5p, hsa-miR-192-3p*, *hsa-miR-4468*, *hsa-miR-4469, hsa-miR-6731-5p*, and *hsa-miR-6835-3p* for miRNA analyses) (Tab. 1, 2 and Suppl. Tab. 1) and subsequently validate their expression stability in exosomes isolated from sera of patients with ovarian cancer and of healthy control individuals. All ten identified candidate reference genes provide better accuracy in terms of stability and variation of expression compared to classical housekeeping genes (Table 1, Fig. 2D and F). Interestingly, if we applied our algorithm to exosome data of each individual cancer type, the predicted candidate reference genes were different from those in the pooled cancer analysis and also varied among cancer subtypes (Suppl Fig. 1).

By employing two different indicators, we define mRNA and miRNA expression stability from two different perspectives: 1) the coefficient of variation "CV" and 2) the "M" value which are both based on the expression abundance of a gene (measured as transcripts per million – TPM) or miRNA (measured as counts per million – CPM). Whereas "CV" measures the variation of a reference gene across all samples, "M" represents the average pairwise variation of a reference gene versus all other reference genes across all samples. Low "CV" and "M" values suggest stable gene expression, and in general, genes with "M" <1.5 can be accepted as reference genes [54]. By requiring a reference gene to have ideally

both, low “CV” and “M” values, we identify *OAZ1* and *hsa-miR-6835-3p* as the most accurate single reference genes based on RNA-Seq datasets (Tab. 1 and 2). We confirm the utility of all candidate reference genes and miRNAs but especially of *OAZ1* and *hsa-miR-6835-3p* by measuring their expression abundance (Fig. 3A and Fig. 4C, D) and stability (Fig. 3B, C, D, E and Fig. 4F, G, I, J) in serum exosomes of an independent cohort of ovarian cancer patients and of healthy control individuals (n=10 each group, Fig. 3). NormFinder [56] is another popular method without requiring priori knowledge of control gene as calibrator to enhance the accuracy, by calculating intra and intergroup variations to evaluate the stability of expression. Here we used NormFinder to evaluate the candidates predicted by geNorm. The results (Suppl. Fig. 5) showed that the most reliable genes *OAZ1* and *hsa-miR-6835-3p* still presented a robust stability measured by NormFinder, with *OAZ1* ranking the 1st in healthy controls and the 2nd in ovarian cancer patients (Suppl. Fig. 5A, B), and with *hsa-miR-6835-3p* ranking the 1st in healthy controls and the 3rd in ovarian cancer (Suppl. Fig. 5C, D). To increase the diagnostic accuracy, some studies suggest the use of a combination of multiple reference genes or miRNAs [52, 54, 57]. We therefore tested various combinations of reference gene candidates, and identified the combinations of *OAZ1*, *SERF2* and *MPP1* for mRNA (Suppl. Fig. 2) and *hsa-miR-6835-3p* and *hsa-miR-4468* for miRNA (Suppl. Fig. 4) providing the highest expression stability across samples and tumor types.

Unlike many previous studies which used an approach to narrow down the number of reference genes from a panel of previously reported candidate genes - usually a list of classical housekeeping genes [37, 45, 58, 59] - our study has the clear advantage of an

294 unbiased and sequencing data-driven approach thus preventing bias from artificial
295 selection. Due to the scarcity of publicly available RNA- and miRNA-Seq exosome
296 datasets with high sequencing quality, the exosome datasets used herein include the
297 RNA-Seq exosome samples with PAAD, CRC and HCC as well as miRNA-Seq exosome
298 samples with HCC, HNSCC, LCA, NBL, OVA, and THCA. Accounting for the limitation
299 that the samples in our current analysis may not fully capture the dynamic expression of
300 genes in exosomes of pan-cancers, our analysis will be continuously updated with the
301 emergence of additional sequencing datasets to refine the robustness of identified candidate
302 reference genes. In this regard, it will be interesting to determine if the reference genes and
303 miRNAs identified here will also proof utility in a true pooled cancer analysis including
304 most or even all of the different cancer types affecting patients.

305 In our analysis, pooled cancer RNA- or miRNA-Seq datasets are from the extracellular
306 vesicles (including exosomes) isolated from the different labs. However, different sized
307 vesicles are derived from different intracellular processes, that affects the numbers and
308 biomolecule contents. Other technical factors (such as exosome isolation and
309 quantification procedures) and biological factors (such as cancer type) can also impact the
310 numbers and composition of exosomes [60-62]. Exosomes isolated from the different
311 sources or by the different isolation methodologies introduce variations in the
312 concentration, purity and size [28, 29]. Moreover, due to the limited knowledge of
313 exosome specific molecular machineries of biogenesis and release, there are the challenges
314 in confirming the biogenesis mechanisms of exosomes. Therefore, future efforts should
315 definitely include standardization (number, volume, etc) and verification of the exosome or

316 extracellular vesicle characteristics, prior to sequencing.

317 We finally confirm previous study results indicating, that housekeeping genes in
318 cancer cell lines and tumor tissue cannot be transferred to the analysis of exosomes and
319 vice versa without further validation[22, 31, 35, 36]. For this, we predicted the top ten
320 candidate reference genes (*ACTB*, *RPS27*, *RPS11*, *RPL13A*, *RPL41*, *RPS14*, *RPL41P1*,
321 *RPS29*, *RPL10* and *NACA*), in cancer tissues with simultaneous high expression abundance
322 (log2(TPM+1)) and stable expression ("CV" and "M") (Suppl. Fig. 6A) , and compared
323 their ranks with those of the top ten predicted reference genes in exosomes and vice versa
324 (Suppl. Fig. 6B). The top ten genes with the high expression stability in tissues as
325 indicated by low "CV" and "M" values were ranked the 120th to 600th in exosomes (Suppl.
326 Fig. 6B, C, D). The results clearly show that there is no overlap between candidate
327 reference genes in exosomes and cancer cell lines/tissue if both, exosome and cancer cell
328 line datasets are not considered together in the initial step of candidate reference prediction
329 (Suppl. Fig. 6B). Although derived from tissue, exosomes deliver specific cargo of protein,
330 miRNA, and small molecules, which is heterogenous between exosome and tissue. The use
331 of tissue-specific reference genes causes quantitative inaccuracy due to the instability of
332 them in exosome studies. Therefore, it is great practical necessity for use exosome-specific
333 reference genes to enhance the quantity accuracy.

334

335 ## Conclusions

336 Our study provides reference genes and miRNAs with high abundance and expression

stability across samples and cancer types for more accurate future qRT-PCR analyses of cancer exosomes. *OAZ1* and *hsa-miR-6835-3p* were identified as the most reliable individual reference genes for mRNA and miRNA quantification, respectively. For superior accuracy, we recommend the use of a combination of reference genes - *OAZ1*/*SERF2*/*MPP1* for mRNA and *hsa-miR-6835-3p*/*hsa-miR-4468-3p* for miRNA analyses. The use of the ideal reference genes is favorable to the accurate quantification of mRNA and miRNA expression fraction within the exosome.

Methods

Data collection

We manually curated RNA- and miRNA-Seq exosome data from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) [63]. The RNA-Seq datasets included 79 serum exosome samples from patients with pancreatic adenocarcinoma (PAAD), colorectal cancer (CRC), hepatocellular carcinoma (HCC) and healthy control individuals (HC) (Suppl. Tab. 2). The miRNA-Seq datasets included 72 serum exosome samples from patients with seven different cancer types (HCC, lung cancer (LCA), ovarian cancer (OVA), head and neck squamous cell carcinoma (HNSCC), neuroblastoma (NBL) and thyroid cancer (THCA)) as well as healthy control individuals (n=31) (Suppl. Tab. 2). For comparison of expression patterns between exosomes and cancer tissues or cell lines, we furthermore analyzed RNA-Seq data from 28 datasets including three cancer types (PAAD, CRC, and HCC) with overall 1,038 tissue and cancer

cell line samples (Suppl. Tab. 2). To validate the reference genes identified by mRNA- and miRNA-Seq analyses, the serum samples were collected from 10 patients diagnosed with ovarian cancer stage I and II according to TNM staging without any treatment and 10 healthy volunteers were recruited in the study.

Processing of raw RNA- and miRNA-Seq data

Raw reads of the RNA-Seq data were trimmed by removing adaptors and low-quality bases using Trimmomatic (version 0.36) [64]. The clean reads were then mapped onto a human reference genome (GRCh38) using HASAT2 software (version 2.1.0) [65]. StringTie (version 1.3.0) [66] was applied to quantify the number of reads that were aligned to the regions of protein-coding RNAs (mRNAs) and annotations of mRNAs in the human genome were retrieved from GENCODE (v29) [67]. For the identification of reference miRNAs, sequencing data were additionally discarded of low-quality reads, adaptor dimers and sequences with lengths < 18 and > 35 nucleotides. The filtered reads were mapped to the human genome by bowtie (version 1.2.1.1) [68] and quantified by featureCounts (version 1.5.3) [69], miRNA annotations were retrieved from miRBase (v22.1) [70]. Expression levels were depicted as transcripts per million for mRNA (TPM) [71] and counts per million (CPM) for miRNA.

Strategy to identify candidate reference genes and miRNAs in cancer exosomes

For individual cancer subtypes, we selected genes or miRNAs as reference candidates if the respective expression level was greater than the average genome-wide expression.

379 Therefore, genes or miRNAs were analyzed using DESeq2 (version 1.22.1) in serum
380 exosomes of cancer patients and of healthy control individuals [72]. DESeq2 provides
381 algorithms for determining differential expression within digital expression datasets using
382 a negative binomial distribution model. In the present study, genes with a *p*-value > 0.1
383 were considered to be not differentially expressed.

384 To determine the expression stability of candidate reference genes and of miRNAs
385 across samples and cancer types, we used two measures: 1) "CV" - the variation of a
386 candidate reference gene or miRNA across all samples - and 2) "M" - the average
387 expression stability value which represents the average pairwise variation between a
388 reference gene and other reference genes across samples calculated by geNorm [54].
389 Lower values of "CV" or "M" indicate higher expression stability.

390 We assumed that there were *m* samples and *n* reference genes. For a given reference
391 gene j in the ith sample, the gene expression level a_{ij} is the normalization expression
392 value (TPM for mRNA, CPM for miRNA) of RNA-/miRNA-Seq data or of the
393 transformed value of cycle threshold numbers (Ct) in qRT-PCR data (Equation 1). Based
394 on the expression levels of the reference gene j across *m* samples (Equation 2), we defined
395 the coefficient of variation "CV" as the ratio of the standard deviation to the mean
396 (Equation 3).

397 $(\forall\, i \in [1,m], \forall\, j \in [1,n])$:

398 $$a_{ij} = 2^{-C_{t_{ij}}} \quad (1)$$

399 $$A_j = \left(a_{1j}, a_{2j}, \ldots a_{mj}\right) = \left(a_{ij}\right)_{i=1\rightarrow m} \quad (2)$$

$$CV_j = \frac{st.dev(A_j)}{mean(A_j)} \quad (3)$$

The second expression stability measure - the average expression stability value "M" - was developed by Vandesompele J *et al.* in the tool geNorm [54]. For any two reference genes j and k, the logarithm of the expression ratio ${a_{ij}}/{a_{ik}}$ for the sample i ($i = 1 \rightarrow m$) forms an array A_{jk} (Equation 4). Based on the pairwise variation V_{jk} defined by the standard deviation of A_{jk} (Equation 5), the average expression stability value M_j for the gene j is the arithmetic mean of all pairwise variation V_{jk} (Equation 6). Usually, a gene with a M less than 1.5 is acceptable as a reference gene.

$(\forall\, j,k \in [1,n]\ and\ j \neq k)$:

$$A_{jk} = \left\{\log_2\left(\frac{a_{1j}}{a_{1k}}\right), \log_2\left(\frac{a_{2j}}{a_{2k}}\right), \ldots \log_2\left(\frac{a_{mj}}{a_{mk}}\right)\right\} = \left\{\log_2\left(\frac{a_{ij}}{a_{ik}}\right)\right\}_{i=1\rightarrow m} \quad (4)$$

$$V_{jk} = st.dev\left(A_{jk}\right) \quad (5)$$

$$M_j = \frac{\sum_{k=1}^{n} V_{jk}}{n-1} \quad (6)$$

Furthermore, multiple genes were combined as candidate references and their stability was measured as the average gene-specific variation AV according to geNorm. For any three reference genes j, k and l, the average gene-specific variation $AV_{j,k,l}$ was calculated as the geometric mean of the three-gene stability value "M" (Equation 7).

$(\forall\, j,k,l \in [1,n]\ and\ j \neq k \neq l)$:

$$AV_{j,k,l} = \sqrt[3]{M_j \cdot M_k \cdot M_l} \quad (7)$$

Isolation of the exosome fraction and sample storage

To validate the reference genes identified by mRNA- and miRNA-Seq analyses, serum exosomes were obtained from ten ovarian cancer patients and ten healthy control individuals. Therefore, blood was drawn and kept at room temperature (15–25°C) for 10 min to 1 h before further processing. The tubes were centrifuged for 10 min at 1,900 x g (3,000 rpm) and 4°C using a swinging bucket rotor. The upper serum phase was transferred into a new canonical tube without disturbing the cell pellet. Subsequently, sera were centrifuged for 15 min at 3,000 x g and 4°C in conical tubes, passed through a 0.8-µm filter and the supernatants were carefully removed without disturbing the pellet and transferred into new tubes. Samples were stored at 2–8°C for immediate processing or kept frozen in aliquots at –65°C to –90°C for long-term storage [29].

Quantitative reverse transcription-PCR (qRT-PCR)

Total RNA (containing the mRNA and miRNA fractions) was isolated from 1 ml serum using the exoRNeasy Serum/Plasma Maxi Kit (Qiagen, Wetzlar, Germany). cDNA was generated with the PrimeScriptRT Master Mix (Perfect Real Time; Takara Bio, Kusatsu, Japan) according to the manufacturer's protocol. Quantitative RT-PCR was performed with ChamQTM Universal SYBR qPCR Master Mix (Vazyme, Nanjing, China) in a final reaction volume of 10 µl using an Applied Biosystems QuantStudio 5 Real-Time PCR Instrument (Thermo Fisher Scientific, Rockford, USA). *18S rRNA* and ce-miR-39-1 served as internal controls for mRNA and miRNA, respectively. Expression levels are depicted as cycle threshold (Ct) value of the candidate gene relative to the Ct value of the housekeeping gene. Data were analyzed with the QuantStudio TM Design & Analysis

442 software. The primer sequences for B2M were 5’-TGTCTTTCAGCAAGGACTGGT-3’
443 and 5’-TGCTTACATGTCTCGATCCCAC-3’, for OAZ1 were
444 5’-GCCAAACGCATTAACTGGCG-3’ and 5’-TGTCCTCGCGGTTCTTGTG-3’, for
445 ITM2B were 5’-TTGCCTCAGTCCTATCTGATTCA-3’ and
446 5’-TCTGCGTTGCAGTTTGTAAGT-3’, for SOD2 were
447 5’-GGAAGCCATCAAACGTGACTT-3’ and 5’-CCCGTTCCTTATTGAAACCAAGC-3’,
448 for PCMTD1 were 5’-TGCATTTGTTGTTGGTAATTGCC-3’ and
449 5’-GTCCAGTTCGCATAATCTGTGT-3’, for ARF1 were
450 5’-ATGGGGAACATCTTCGCCAAC-3’ and 5’-GTGGTCACGATCTCACCCAG-3’, for
451 MPP1 were 5’-GTCAGCTCCTAGCGAAGCC-3’ and
452 5’-GCCGAACGACTTCCTCGTAG-3’, for WIPF1 were
453 5’-AGCCGCTGCGCGATTTAT-3’ and 5’-TCCCAGCCTGCTCTGTCTTA-3’, for
454 SERF2 were 5’-CCGCAAGCAGAGGGACTC-3’ and
455 5’-AGCACTACAGGAGGAAACGC-3’, for H3F3AP4 were
456 5’-CAGCTATCGGTGCTTTGCAG-3’ and 5’-AGCACGTTCTCCACGTATGC-3’, for
457 18s were 5’-CTTCCACAGGAGGCCTACAC-3’ and
458 5’-CTTCGGCCCACACCCTTAAT-3’.

459

460 **Statistical analysis**

461 The Wilcoxon rank-sum test was used to compare expression stability between classical
462 housekeeping and exosomal candidate reference genes as measured by “CV” and “M”.

Candidate reference genes were sorted according to their “CV” and “M” values from low (higher expression stability across samples) to high (lower expression stability across samples) and assigned a rank, and the best candidate gene or miRNA for validation was determined as the one with the lowest sum of these two ranks. All statistical analyses were executed in R.

List of abbreviations

EVs: Extracellular vesicles

mRNA: Messenge RNA

miRNA: MicroRNA

RNA: Ribonucleic acid

RNA-Seq: RNA-sequencing

miRNA-Seq：MicroRNA-sequencing

qRT-PCR: Quantitative real-time polymerase chain reaction

TPM: Transcripts per million

CPM: Counts per million

Ct: Cycle threshold numbers

CV: Coefficient of variation

PAAD: Pancreatic adenocarcinoma

CRC: Colorectal cancer

HCC: Hepatocellular carcinoma

HC: Healthy control individuals

LCA: Lung cancer

OVA: Ovarian cancer

HNSCC: Head and neck squamous cell carcinoma

NBL: Neuroblastoma

THCA: Thyroid cancer

B2M: beta-2-microglobulin

ARF1: ADP-ribosylation factor 1

H3F3AP4: H3 histone pseudogene 6

ITM2B: integral membrane protein 2B

MPP1: membrane palmitoylated protein 1

OAZ1: ornithine decarboxylase antizyme 1

PCMTD1: protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 1

SOD2: superoxide dismutase 2

SERF2: small EDRK-rich factor 2

WIPF1: WAS/WASL Interacting Protein Family Member 1

ACTB: beta-actin

RPL13A: ribosomal protein L13A

YWHAZ: tyrosine 3-monooxygenase/tryptophane 5-monooxygenase activation protein zeta

GAPDH: glyceraldehyde-3-phosphate dehydrogenase

VIM: vimentin

PPIA: peptidylprolyl isomerase A

ALDOA: aldolase A

UBC: ubiquitin C

FTL: ferritin light chain,

FYB1: FYN binding protein 1

ARPC2: actin related protein 2/3 complex subunit 2

NCOA4: nuclear receptor coactivator 4

HCLS1: hematopoietic cell-specific Lyn substrate 1

TYROBP: transmembrane immune signaling adaptor TYROBP

RPL41: ribosomal protein L41

SNCA: synuclein alpha

RPS9: ribosomal protein S9

BTF3: basic transcription factor 3

ADIPOR1: adiponectin receptor 1

HEMGN: hemogen

CD74: CD74 molecule

DDX5: DEAD-box helicase 5

HCLS1: hematopoietic cell-specific Lyn substrate 1

LSP: lymphocyte specific protein 1

RPL9: ribosomal protein L9

Ethics approval and consent to participate

This study has been approved by the ethics committee of Shanghai First Maternity and Infant Hospital, Tongji University School of Medicine (Ethics code: KS1748). All of the participants signed an informed consent and the investigators obtained the written informed consent prior to entry into the study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study are available in the Gene Expression Omnibus (GEO), [https://www.ncbi.nlm.nih.gov/geo/]. The datasets information is included in the supplementary materials.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the National Key Research and Development Program (2017YFC0908500 to HW), the National Natural Science Foundation of China (31771469 and 31571363 to HW, and 81772762 to XS), the Shanghai Natural Science Foundation of China (19ZR1461600 to AL) and a Mildred-Scheel post-doctoral fellowship of the German Cancer Aid Foundation (70111755 to JK). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

HW conceived the hypothesis. YD and HW designed and performed the data analysis. YC and SX performed the experimental validation. HW, YD, SX and JK interpreted the results and wrote the manuscript. AL provided helpful suggestions for the data analysis.

Acknowledgements

We thank all individuals who participated in this study.

References

1. Thery C, Zitvogel L, Amigorena S: **Exosomes: composition, biogenesis and function**. *Nature reviews Immunology* 2002, **2**(8):569-579.
2. Ruivo CF, Adem B, Silva M, Melo SA: **The Biology of Cancer Exosomes: Insights and New Perspectives**. *Cancer research* 2017, **77**(23):6480-6488.
3. Valadi H, Ekstrom K, Bossios A, Sjostrand M, Lee JJ, Lotvall JO: **Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells**. *Nature Cell Biology* 2007, **9**(6):654.
4. Yáñez-Mó M, Siljander PRM, Andreu Z, Bedina Zavec A, Borràs FE, Buzas EI, Buzas K, Casal E, Cappello F, Carvalho J *et al*: **Biological properties of extracellular vesicles and their physiological functions**. *Journal of extracellular vesicles* 2015, **4**(1):27066.
5. Jeppesen DK, Fenix AM, Franklin JL, Higginbotham JN, Zhang Q, Zimmerman LJ, Liebler DC, Ping J, Liu Q, Evans R *et al*: **Reassessment of Exosome Composition**. *Cell* 2019, **177**(2):428-445.e418.
6. Redis RS, Calin S, Yang Y, You MJ, Calin GA: **Cell-to-cell miRNA transfer: from body homeostasis to therapy**. *Pharmacology & therapeutics* 2012, **136**(2):169-174.
7. Tkach M, Thery C: **Communication by Extracellular Vesicles: Where We Are and Where We Need to Go**. *Cell* 2016, **164**(6):1226-1232.
8. Kalluri R: **The biology and function of exosomes in cancer**. *The Journal of clinical investigation* 2016, **126**(4):1208-1215.
9. Zhao H, Yang L, Baddour J, Achreja A, Bernard V, Moss T, Marini JC, Tudawe T, Seviour EG, San Lucas FA *et al*: **Tumor microenvironment derived exosomes pleiotropically modulate cancer cell metabolism**. *Elife* 2016, **5**:e10250-e10250.
10. Alipoor SD, Mortaz E, Varahram M, Movassaghi M, Kraneveld AD, Garssen J, Adcock IM: **The Potential Biomarkers and Immunological Effects of Tumor-Derived Exosomes in Lung Cancer**. *Frontiers in immunology* 2018, **9**:819.
11. Maia J, Caja S, Strano Moraes MC, Couto N, Costa-Silva B: **Exosome-Based Cell-Cell Communication in the Tumor Microenvironment**. *Frontiers in cell and developmental biology* 2018, **6**:18.
12. Xu R, Rai A, Chen M, Suwakulsiri W, Greening DW, Simpson RJ: **Extracellular vesicles in cancer — implications for future improvements in cancer care**. *Nature Reviews Clinical Oncology* 2018, **15**(Suppl):617-638.
13. Salehi M, Sharifi M: **Exosomal miRNAs as novel cancer biomarkers: Challenges and opportunities**. *Journal of cellular physiology* 2018, **233**(9):6370-6380.
14. Vaidyanathan R, Soon RH, Zhang P, Jiang K, Lim CT: **Cancer diagnosis: from tumor to liquid biopsy and beyond**. *Lab on a chip* 2018, **19**(1):11-34.
15. Bernard V, Kim DU, San Lucas FA, Castillo J, Allenson K, Mulu FC, Stephens BM, Huang J, Semaan A, Guerrero PA *et al*: **Circulating Nucleic Acids Are Associated With Outcomes of Patients With Pancreatic Cancer**. *Gastroenterology* 2019, **156**(1):108-118 e104.
16. Kalishwaralal K, Kwon WY, Park KS: **Exosomes for Non-Invasive Cancer Monitoring**. *Biotechnology journal* 2019, **14**(1):e1800430.
17. Kalluri R, LeBleu VS: **The biology, function, and biomedical applications of exosomes**. *Science* 2020, **367**(6478).

18. Sheridan C: **Exosome cancer diagnostic reaches market**. *Nature Biotechnology* 2016, **34**(4):359-360.
19. McKiernan J, Donovan MJ, O' Neill V, Bentink S, Noerholm M, Belzer S, Skog J, Kattan MW, Partin A, Andriole G *et al*: **A Novel Urine Exosome Gene Expression Assay to Predict High-grade Prostate Cancer at Initial Biopsy**. *JAMA Oncology* 2016, **2**(7):882-889.
20. Hoshino A, Kim HS, Bojmar L, Gyan KE, Cioffi M, Hernandez J, Zambirinis CP, Rodrigues G, Molina H, Heissel S *et al*: **Extracellular Vesicle and Particle Biomarkers Define Multiple Human Cancers**. *Cell* 2020, **182**(4):1044-1061 e1018.
21. VanGuilder HD, Vrana KE, Freeman WM: **Twenty-five years of quantitative PCR for gene expression analysis**. *Biotechniques* 2008, **44**(5):619-626.
22. Huggett J, Dheda K, Bustin S, Zumla A: **Real-time RT-PCR normalisation; strategies and considerations**. *Genes & Immunity* 2005, **6**(4):279-284.
23. Derveaux S, Vandesompele J, Hellemans J: **How to do successful gene expression analysis using real-time PCR**. *Methods* 2010, **50**(4):227-230.
24. S A B: **Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays**. *Journal of molecular endocrinology* 2000, **2**(25):169-193.
25. Ruijter JM, Ramakers C, Hoogaars WMH, Karlen Y, Bakker O, van den Hoff MJB, Moorman AFM: **Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data**. *Nucleic Acids Research* 2009, **37**(6):e45-e45.
26. Schmittgen TD, Zakrajsek BA: **Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR**. *J Biochem Biophys Methods* 2000, **46**(1-2):69-81.
27. Ludwig N, Whiteside TL, Reichert TE: **Challenges in Exosome Isolation and Analysis in Health and Disease**. *Int J Mol Sci* 2019, **20**(19).
28. Konoshenko MY, Lekchnov EA, Vlassov AV, Laktionov PP: **Isolation of Extracellular Vesicles: General Methodologies and Latest Trends**. *Biomed Res Int* 2018, **2018**:8545347.
29. Tang YT, Huang YY, Zheng L, Qin SH, Xu XP, An TX, Xu Y, Wu YS, Hu XM, Ping BH *et al*: **Comparison of isolation methods of exosomes and exosomal RNA from cell culture medium and serum**. *Int J Mol Med* 2017, **40**(3):834-844.
30. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL *et al*: **The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments**. *Clinical chemistry* 2009, **55**(4):611-622.
31. Kozera B, Rapacz M: **Reference genes in real-time PCR**. *Journal of Applied Genetics* 2013, **54**(4):391-406.
32. Eisenberg E, Levanon EY: **Human housekeeping genes are compact**. *Trends in Genetics* 2003, **19**(7):362-365.
33. Noerholm M, Balaj L, Limperg T, Salehi A, Zhu LD, Hochberg FH, Breakefield XO, Carter BS, Skog J: **RNA expression patterns in serum microvesicles from patients with glioblastoma multiforme and controls**. *BMC Cancer* 2012, **12**(1):22.
34. Thind A, Wilson C: **Exosomal miRNAs as cancer biomarkers and therapeutic targets**. *J Extracell Vesicles* 2016, **5**:31292.
35. Ohl F, Jung M, Xu C, Stephan C, Rabien A, Burkhardt M, Nitsche A, Kristiansen G, Loening SA, Radonić A *et al*: **Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization?** *Journal of Molecular Medicine* 2005,

83(12):1014-1024.

36. Goidin D, Mamessier A, Staquet M-J, Schmitt D, Berthier-Vergnes O: **Ribosomal 18S RNA Prevails over Glyceraldehyde-3-Phosphate Dehydrogenase and β-Actin Genes as Internal Standard for Quantitative Comparison of mRNA Levels in Invasive and Noninvasive Human Melanoma Cell Subpopulations**. *Analytical Biochemistry* 2001, **295**(1):17-21.

37. Lamba V, Ghodke-Puranik Y, Guan W, Lamba JK: **Identification of suitable reference genes for hepatic microRNA quantitation**. *BMC Research Notes* 2014, **7**(1):129.

38. Ramezani A, Devaney JM, Cohen S, Wing MR, Scott R, Knoblach S, Singhal R, Howard L, Kopp JB, Raj DS: **Circulating and urinary microRNA profile in focal segmental glomerulosclerosis: a pilot study**. *European Journal of Clinical Investigation* 2015, **45**(4):394-404.

39. Solé C, Cortés-Hernández J, Felip ML, Vidal M, Ordi-Ros J: **miR-29c in urinary exosomes as predictor of early renal fibrosis in lupus nephritis**. *Nephrology Dialysis Transplantation* 2015, **30**(9):1488-1496.

40. Zheng G, Wang H, Zhang X, Yang Y, Wang L, Du L, Li W, Li J, Qu A, Liu Y *et al*: **Identification and Validation of Reference Genes for qPCR Detection of Serum microRNAs in Colorectal Adenocarcinoma Patients**. *PLOS ONE* 2013, **8**(12):e83025.

41. Wang K, Yuan Y, Cho J-H, McClarty S, Baxter D, Galas DJ: **Comparing the MicroRNA Spectrum between Serum and Plasma**. *PLOS ONE* 2012, **7**(7):e41561.

42. Benz F, Roderburg C, Vargas Cardenas D, Vucur M, Gautheron J, Koch A, Zimmermann H, Janssen J, Nieuwenhuijsen L, Luedde M *et al*: **U6 is unsuitable for normalization of serum miRNA levels in patients with sepsis or liver fibrosis**. *Experimental & Molecular Medicine* 2013, **45**(9):e42-e42.

43. Xiang M, Zeng Y, Yang R, Xu H, Chen Z, Zhong J, Xie H, Xu Y, Zeng X: **U6 is not a suitable endogenous control for the quantification of circulating microRNAs**. *Biochemical and Biophysical Research Communications* 2014, **454**(1):210-214.

44. Lange T, Stracke S, Rettig R, Lendeckel U, Kuhn J, Schlüter R, Rippe V, Endlich K, Endlich N: **Identification of miR-16 as an endogenous reference gene for the normalization of urinary exosomal miRNA expression data from CKD patients**. *PLOS ONE* 2017, **12**(8):e0183435.

45. Ragni E, Perucca Orfei C, De Luca P, Colombini A, Viganò M, Lugano G, Bollati V, de Girolamo L: **Identification of miRNA Reference Genes in Extracellular Vesicles from Adipose Derived Mesenchymal Stem Cells for Studying Osteoarthritis**. *International Journal of Molecular Sciences* 2019, **20**(5):1108.

46. Francis J, Rea G, Stephanie N, Sheri N, André F, Hacker NF, Viola HS: **Careful selection of reference genes is required for reliable performance of RT-qPCR in human normal and cancer cell lines**. *Plos One* 2013, **8**(3):e59180.

47. Mathieu M, Martin-Jaular L, Lavieu G, Thery C: **Specificities of secretion and uptake of exosomes and other extracellular vesicles for cell-to-cell communication**. *Nat Cell Biol* 2019, **21**(1):9-17.

48. Okoye Isobel S, Coomes Stephanie M, Pelly Victoria S, Czieso S, Papayannopoulos V, Tolmachova T, Seabra Miguel C, Wilson Mark S: **MicroRNA-Containing T-Regulatory-Cell-Derived Exosomes Suppress Pathogenic T Helper 1 Cells**. *Immunity* 2014, **41**(1):89-103.

49. Skog J, Würdinger T, van Rijn S, Meijer DH, Gainche L, Curry WT, Carter BS, Krichevsky AM,

Breakefield XO: **Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers**. *Nature Cell Biology* 2008, **10**(12):1470-1476.

50. Yan W, Wu X, Zhou W, Fong MY, Cao M, Liu J, Liu X, Chen C-H, Fadare O, Pizzo DP *et al*: **Cancer-cell-secreted exosomal miR-105 promotes tumour growth through the MYC-dependent metabolic reprogramming of stromal cells**. *Nature Cell Biology* 2018, **20**(5):597-609.

51. Huang X, Yuan T, Tschannen M, Sun Z, Jacob H, Du M, Liang M, Dittmar RL, Liu Y, Liang M *et al*: **Characterization of human plasma-derived exosomal RNAs by deep sequencing**. *BMC Genomics* 2013, **14**(1):319.

52. Madadi S, Schwarzenbach H, Lorenzen J, Soleimani M: **MicroRNA expression studies: challenge of selecting reliable reference controls for data normalization**. *Cellular and Molecular Life Sciences* 2019, **76**(18):3497-3514.

53. Geeurickx E, Hendrix A: **Targets, pitfalls and reference materials for liquid biopsy tests in cancer diagnostics**. *Molecular Aspects of Medicine* 2019.

54. Vandesompele J, Preter KD, Pattyn F, Poppe B, Roy NV, Paepe AD, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes**. *Genome Biology* 2002, **3**(7):research0034.0031.

55. Haller F, Kulle B, Schwager S, Gunawan B, Heydebreck Av, Sültmann H, Füzesi L: **Equivalence test in quantitative reverse transcription polymerase chain reaction: confirmation of reference genes suitable for normalization**. *Analytical Biochemistry* 2004, **335**(1):1-9.

56. Andersen CL, Jensen JL, Orntoft TF: **Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets**. *Cancer research* 2004, **64**(15):5245-5250.

57. Nicot N, Hausman J-F, Hoffmann L, Evers D: **Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress**. *Journal of Experimental Botany* 2005, **56**(421):2907-2914.

58. Li Y, Xiang GM, Liu LL, Liu C, Liu F, Jiang DN, Pu XY: **Assessment of endogenous reference gene suitability for serum exosomal microRNA expression analysis in liver carcinoma resection studies**. *Molecular medicine reports* 2015, **12**(3):4683-4691.

59. Song J, Bai Z, Han W, Zhang J, Meng H, Bi J, Ma X, Han S, Zhang Z: **Identification of Suitable Reference Genes for qPCR Analysis of Serum microRNA in Gastric Cancer Patients**. *Digestive Diseases and Sciences* 2012, **57**(4):897-904.

60. Thery C, Witwer KW, Aikawa E, Alcaraz MJ, Anderson JD, Andriantsitohaina R, Antoniou A, Arab T, Archer F, Atkin-Smith GK *et al*: **Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines**. *J Extracell Vesicles* 2018, **7**(1):1535750.

61. Royo F, Thery C, Falcon-Perez JM, Nieuwland R, Witwer KW: **Methods for Separation and Characterization of Extracellular Vesicles: Results of a Worldwide Survey Performed by the ISEV Rigor and Standardization Subcommittee**. *Cells* 2020, **9**(9).

62. Gurunathan S, Kang MH, Jeyaraj M, Qasim M, Kim JH: **Review of the Isolation, Characterization, Biological Function, and Multifarious Therapeutic Approaches of Exosomes**. *Cells* 2019, **8**(4).

63. Tanya B, Troup DB, Wilhite SE, Pierre L, Carlos E, Kim IF, Maxim T, Marshall KA, Phillippy KH, Sherman PM: **NCBI GEO: archive for functional genomics data sets--10 years on**. *Nucleic Acids Research* 2013, **39**(Database issue):1005-1010.
64. Bolger AM, Marc L, Bjoern U: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
65. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nature methods* 2015, **12**(4):357-360.
66. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads**. *Nat Biotechnol* 2015, **33**(3):290-295.
67. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome Research* 2012, **22**(9):1760-1774.
68. Langmead B, Pop M: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology* 2009, **10**(3):R25.
69. Yang L, Smyth GK, Wei S: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics* 2014, **30**(7):923-930.
70. Ana K, Sam GJ: **miRBase: integrating microRNA annotation and deep-sequencing data**. *Nucleic Acids Research* 2011, **39**(Database issue):D152.
71. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**(4):493-500.
72. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biology* 2010, **11**(10):R106.

Figure legends

Figure 1 Scatterplots of predicted candidate reference genes for serum exosomes using RNA-Seq data

Expression levels of candidate reference genes in serum exosomes are depicted for pooled cancer samples (PAAD, CRC, HCC) (**A**), for pancreatic adenocarcinoma (PAAD) (**B**), colorectal cancer (CRC) (**C**) and hepatocellular carcinoma (HCC) (**D**) samples and compared to serum exosomes of healthy control individuals. Expression values are shown as the logarithm of transcripts per million (TPM) ($\log_2$(TPM+1)). Red dots represent candidate reference genes and grey dots genome-wide genes.

Figure 2 Gene expression levels and stability of candidate reference genes for exosomes predicted with RNA-Seq data

Expression levels of ten candidate genes sorted by their respective expression levels (**A**). Expression levels of ten candidate reference genes (blue bars) compared with those of nine commonly used housekeeping genes (green bars) (**B**). Expression stability of candidate reference genes as measured by the coefficient of variation ("CV") (**C**). Comparison of "CV" values between candidate reference genes and classical housekeeping genes (*p=8.10e-04*, Wilcoxon rank-sum test) (**D**). Expression stability of candidate reference genes as measured by the average expression stability value ("M") (**E**). Comparison of "M" values between candidate reference genes and classical housekeeping genes (*p=0.0279*, Wilcoxon rank-sum test) (**F**).

Figure 3 Experimental validation of candidate reference genes in exosomes of patients with ovarian cancer and healthy control individuals

Expression levels (Ct values) of candidate reference genes in exosomes of ovarian cancer patients (red bars) and healthy control individuals (blue bars) relative to *18S rRNA* (**A**). Expression stability of the candidate reference genes in serum exosomes of ovarian cancer patients (**B**) and healthy control individuals (**C**) as measured by the "CV" indicator. Expression stability of the candidate reference genes in serum exosomes of ovarian cancer patients (**D**) and healthy control individuals (**E**) as measured by the "M" indicator.

Figure 4 Identification and validation of candidate reference miRNAs predicted in exosomes of ovarian cancer patients

Scatterplot of candidate reference miRNA expression levels in pooled cancer samples (HCC, HNSCC, LCA, NBL, OVA, and THCA) and healthy control individuals. Expression values are shown as the logarithm of counts per million (CPM) ($\log_2$(CPM+1)). The red dots represent candidate reference miRNAs, grey dots genome-wide miRNAs (**A**). Expression levels of six candidate reference miRNAs in exosomes of pooled cancer (**B**), ovarian cancer patients (relative to ce-miR-39-1, n=10) (**C**) and healthy control individuals (relative to ce-miR-39-1, n=10) (**D**). Expression stability of candidate reference miRNAs in exosomes of pooled cancer (**E**), ovarian cancer patients (**F**) and healthy control individuals (**G**) as measured by the "CV". Expression stability of six candidate reference miRNAs in exosomes of pooled cancer (**H**), ovarian cancer patients (**I**) and healthy control individuals (**J**) as measured by the "M" indicator.

Tables

Table 1 Candidate reference genes (n=10) ranked in order of their expression level and expression stability

Table 2 Candidate reference miRNAs (n=6) ranked in order of their expression level

and expression stability

Additional files

Figure S1 Gene expression level and stability of candidate reference genes in PAAD, HCC and CRC exosomes predicted using RNA-Seq data A, Expression levels of ten candidate genes in PAAD (**A**), HCC (**B**) and CRC (**C**) exosomes sorted by their expression. Expression stability of ten candidate reference genes in exosomes of patients with PAAD (**D**), HCC (**E**) and CRC (**F**) measured by the "CV". Expression stability of ten candidate reference genes in exosomes of PAAD (**D**), HCC (**E**) and CRC (**F**) patients as measured by the "M" indicator. Expression levels are given as the $\log_2$(TPM+1) (TPM – transcripts per million). The TPMs of each candidate gene were used to determine "CV" and "M" indicators.

Figure S2 Evaluation of expression stability of combinations of three reference genes in pooled exosomes of cancer patients with PAAD, CRC and HCC (A, RNA-Seq data) and with ovarian cancer (B, qRT-PCR data) The expression stability of respective combinations was measured as the average gene-specific variation calculated with the geNorm algorithm based on transcripts per million (TPM) (**A**) or cycle threshold (Ct) values (**B**). Three combinations according to their expression stability ranking from Table 1 were evaluated: 1) genes 1-3 (*OAZ1*, *SERF2*, *MPP1*); 2) genes 4-6 (*H3F3AP4*, *WIPF1*, *PCMTD1*); and 3) genes 8-10 (*SOD2*, *B2M*, *ITM2B*).

Figure S3 Scatterplots of expression levels of six candidate reference miRNAs (red dots) in serum exosomes of patients with HCC (A), HNSCC (B), LCA (C), NBL (D), OVA (E) and THCA (F) compared to exosomes of healthy control individuals The expression values are depicted as: $\log_2$(CPM+1) (CPM – counts per million). Grey dots indicate genome-wide miRNAs.

Figure S4 Evaluation of expression stability of combinations of two reference miRNA candidates in pooled exosomes of cancer patients with different tumor types (A, miRNA-Seq data) or with ovarian cancer (B, qPCR data) The expression stability of miRNA combinations was measured as the average miRNA-specific variation, which was calculated by the geNorm algorithm based on counts per million (CPM) (**A**) or cycle threshold (Ct) values (**B**). Three combinations were considered according to the miRNA expression stability ranks shown in Table 2: miRNAs 1-2 (*miR-4468* and *miR-6835-3p*), miRNAs 3-4 (*miR-192-3p* and *miR-125a-5p)*, and miRNAs 5-6 (*miR-4469* and *miR-6731-5p*).

Figure S5 Validation of candidate reference genes and miRNAs predicted in exosomes of ovarian cancer patients and healthy controls by NormFinder Expression stability of candidate reference genes as measured by the NormFinder stability value in exosomes of ovarian cancer patients (**A**) and healthy control individuals (B). Expression

stability of six candidate reference miRNAs in exosomes of ovarian cancer patients (**C**) and healthy control individuals (**D**) as measured by the NormFinder stability value.

Figure S6 Analysis of candidate reference genes predicted in cancer tissues Expression levels of the ten top candidates in pooled cancer tissue samples calculated using RNA-Seq data. Expression levels are given as log2(TPM+1; TPM = transcript per million) (A). The top ten predicted candidate reference genes for exosomes were compared with the respective ranking in cancer tissues and vice versa (B). Expression stability of the ten top candidate reference genes in tumor tissues as measured by "CV" (C) and "M" indicators (D).

Table S1 List of candidate reference genes (n=10) and miRNAs (n=6) identified by RNA-Seq and quantitative real-time PCR analyses

Table S2 Detailed information for the RNA-Seq datasets in the GEO database

Figures

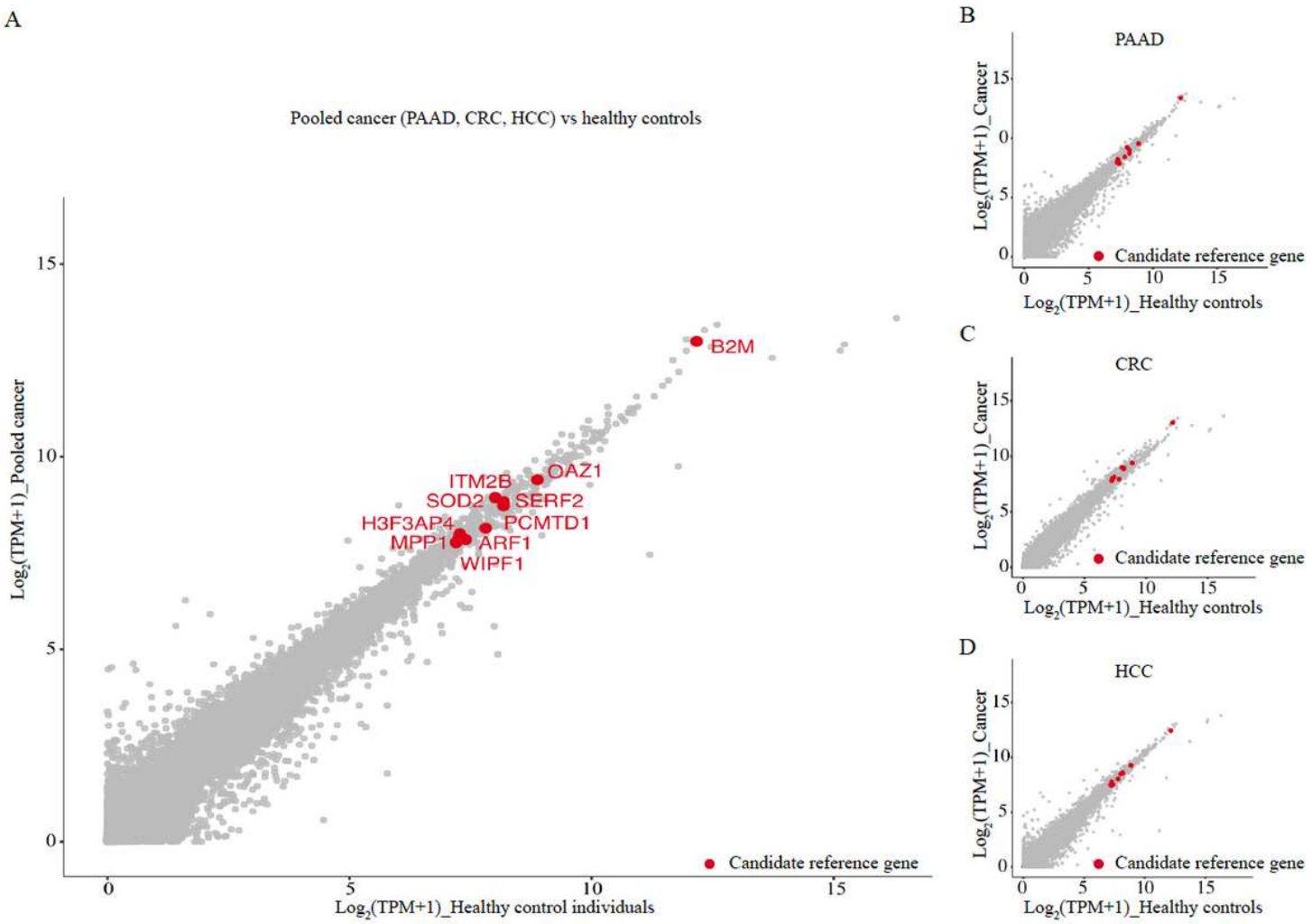

Figure 1

Scatterplots of predicted candidate reference genes for serum exosomes using RNA-Seq data Expression levels of candidate reference genes in serum exosomes are depicted for pooled cancer samples (PAAD, CRC, HCC) (A), for pancreatic adenocarcinoma (PAAD) (B), colorectal cancer (CRC) (C) and hepatocellular carcinoma (HCC) (D) samples and compared to serum exosomes of healthy control individuals. Expression values are shown as the logarithm of transcripts per million (TPM) (log2(TPM+1)). Red dots represent candidate reference genes and grey dots genome-wide genes.

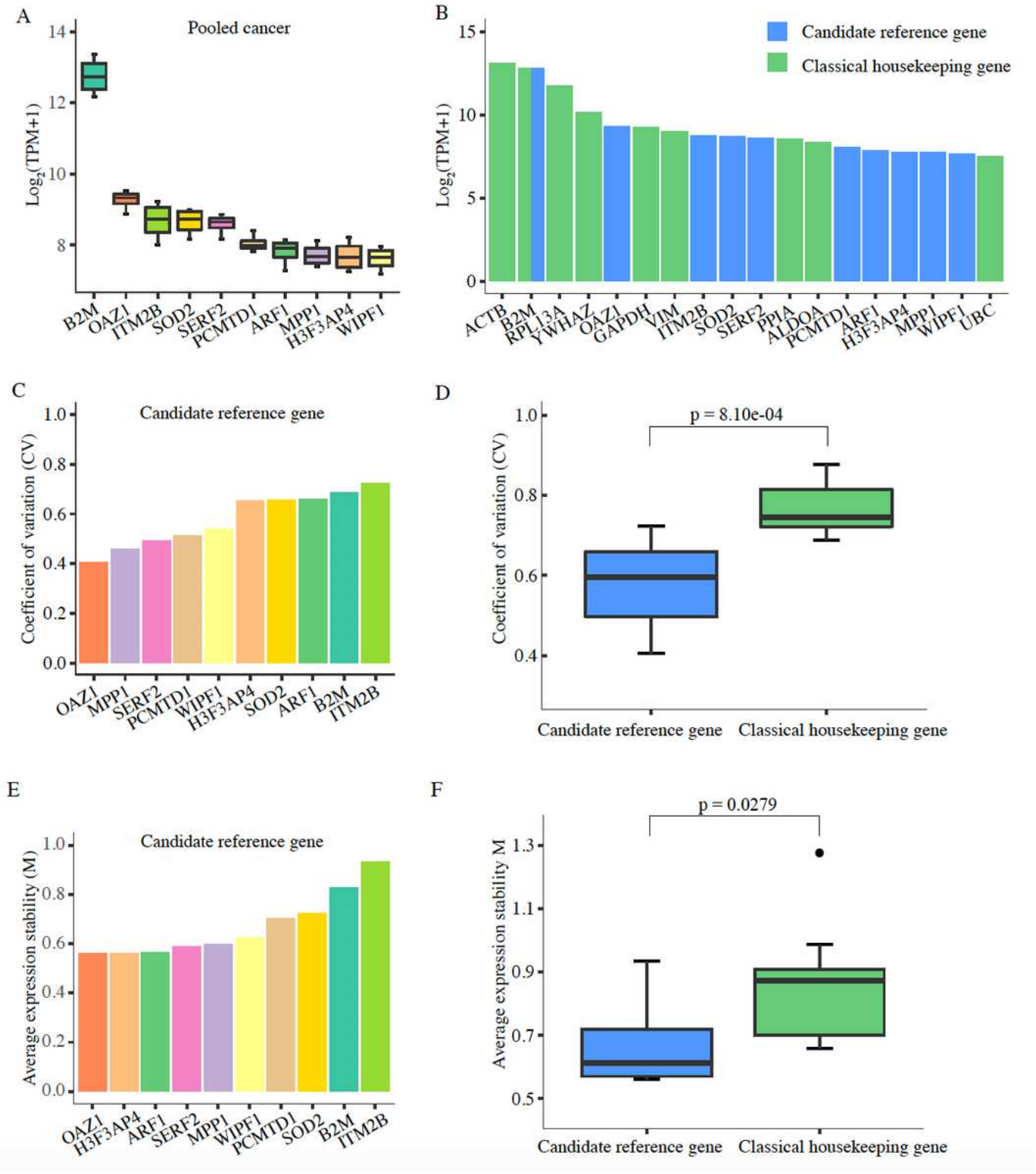

Figure 2

Gene expression levels and stability of candidate reference genes for exosomes predicted with RNA-Seq data Expression levels of ten candidate genes sorted by their respective expression levels (A). Expression levels of ten candidate reference genes (blue bars) compared with those of nine commonly used housekeeping genes (green bars) (B). Expression stability of candidate reference genes as measured by the coefficient of variation (“CV”) (C). Comparison of “CV” values between candidate reference genes and

classical housekeeping genes (p=8.10e-04, Wilcoxon rank-sum test) (D). Expression stability of candidate reference genes as measured by the average expression stability value (“M”) (E). Comparison of “M” values between candidate reference genes and classical housekeeping genes (p=0.0279, Wilcoxon rank-sum test) (F).

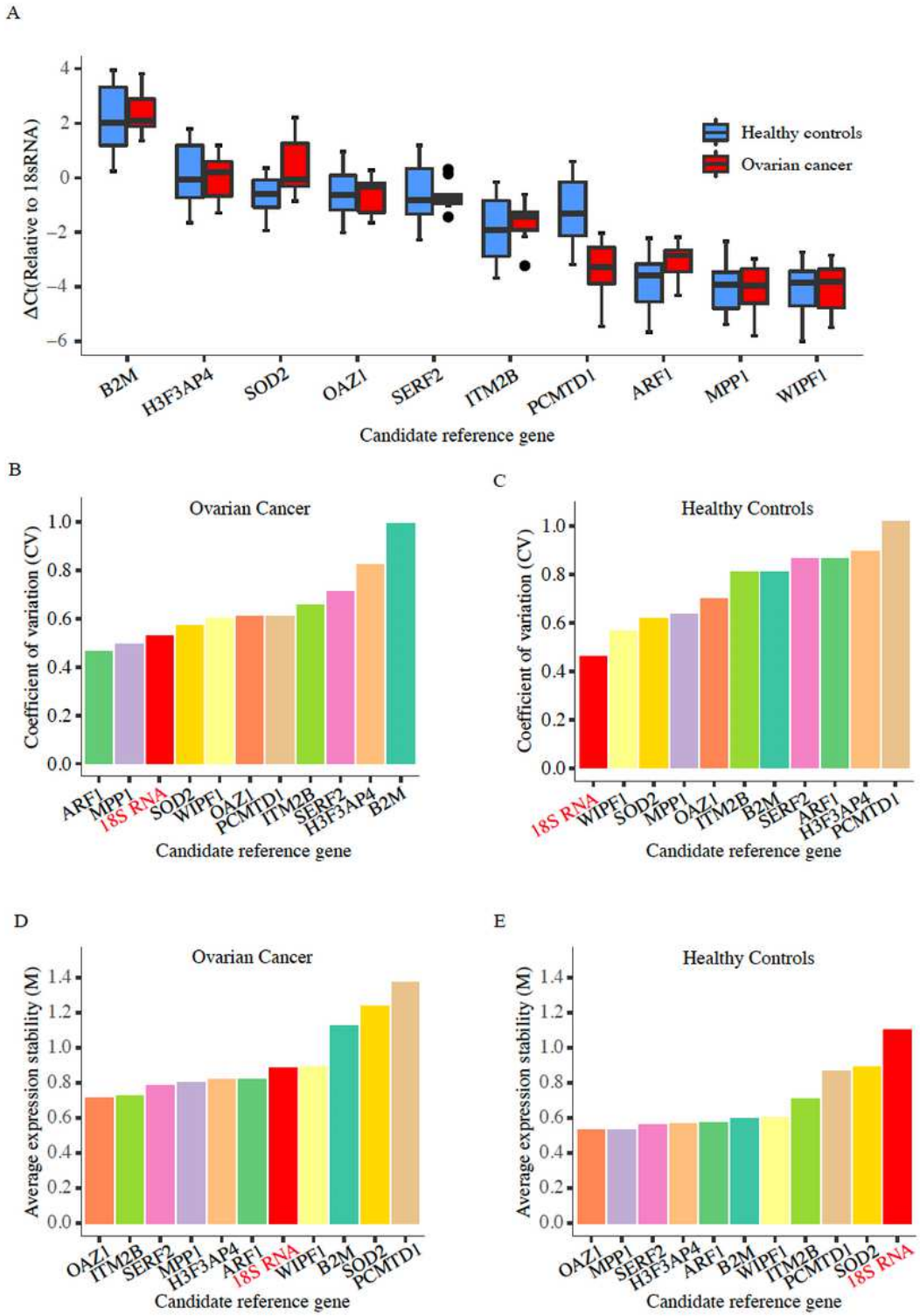

Figure 3

Experimental validation of candidate reference genes in exosomes of patients with ovarian cancer and healthy control individuals Expression levels (Ct values) of candidate reference genes in exosomes of ovarian cancer patients (red bars) and healthy control individuals (blue bars) relative to 18S rRNA (A). Expression stability of the candidate reference genes in serum exosomes of ovarian cancer patients (B) and healthy control individuals (C) as measured by the “CV” indicator. Expression stability of the candidate reference genes in serum exosomes of ovarian cancer patients (D) and healthy control individuals (E) as measured by the “M” indicator.

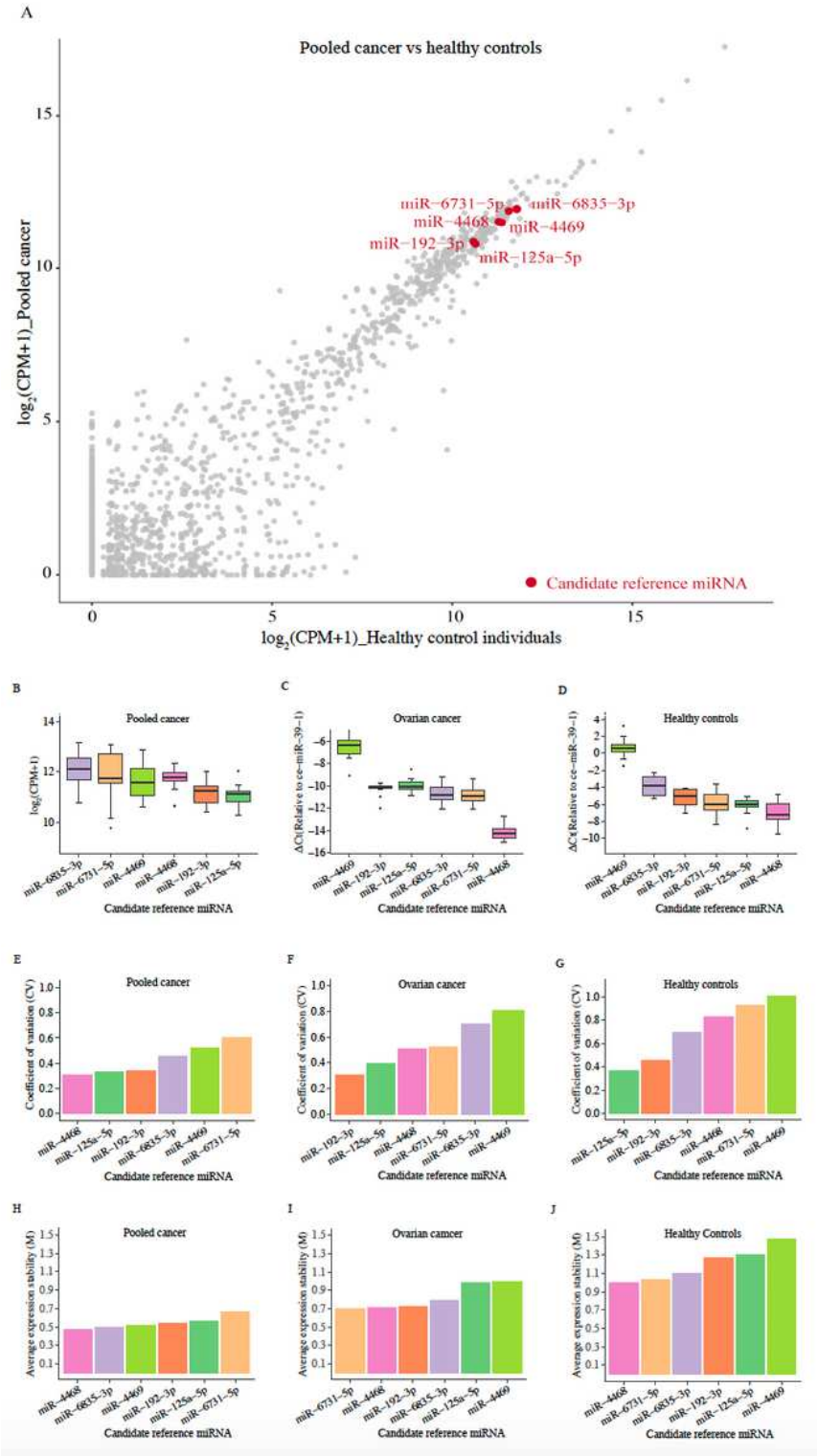

Figure 4

Identification and validation of candidate reference miRNAs predicted in exosomes of ovarian cancer patients Scatterplot of candidate reference miRNA expression levels in pooled cancer samples (HCC, HNSCC, LCA, NBL, OVA, and THCA) and healthy control individuals. Expression values are shown as the logarithm of counts per million (CPM) (log2(CPM+1)). The red dots represent candidate reference miRNAs, grey dots genome-wide miRNAs (A). Expression levels of six candidate reference miRNAs in exosomes of pooled cancer (B), ovarian cancer patients (relative to ce-miR-39-1, n=10) (C) and healthy control individuals (relative to ce-miR-39-1, n=10) (D). Expression stability of candidate reference miRNAs in exosomes of pooled cancer (E), ovarian cancer patients (F) and healthy control individuals (G) as measured by the “CV”. Expression stability of six candidate reference miRNAs in exosomes of pooled cancer (H), ovarian cancer patients (I) and healthy control individuals (J) as measured by the “M” indicator.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableSupplementary1.docx
- TableSupplementary2.docx
- SupplementaryFigure1.png
- SupplementaryFigure2.png
- SupplementaryFigure3.png
- SupplementaryFigure4.png
- SupplementaryFigure5.png
- SupplementaryFigure6.png