

UFLC-derived CSF extracellular vesicle origin and proteome

*AUTHOR LIST: Alexander G. Thompson^{1‡}, Elizabeth Gray^{1‡}, Imre Mager², Roman Fischer³,
Marie-Laëtitia Thézénas³, Philip D Charles³, Kevin Talbot¹, Samir El Andaloussi^{2,4}, Benedikt M.
Kessler³, Matthew Wood^{2*}, Martin R. Turner^{1*}*

Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom¹;

Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United
Kingdom²;

Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford,
United Kingdom³;

Department of Laboratory Medicine, Karolinska Institutet, Stockholm, Sweden⁴

‡ these authors contributed equally

Corresponding authors

*Correspondence:	Professor Martin Turner	Professor Matthew Wood
	West Wing Level 6	Le Gros Clark Building
	John Radcliffe Hospital	South Parks Road
	Oxford OX3 9DU	Oxford OX1 3QX
	Tel: +44 (0)1865 223380	Tel: +44 (0)1865 282840
	martin.turner@ndcn.ox.ac.uk matthew.wood@dpag.ox.ac.uk	

Abbreviations

EV – extracellular vesicle; CSF – cerebrospinal fluid; UC – ultracentrifugation; UFLC – ultrafiltration liquid chromatography; Nano-UPLC-MS/MS – nano ultrahigh performance LC-MS/MS; NTA – nanoparticle tracking analysis; TEM – transmission electron microscopy; OR – odds ratio.

Word counts: Abstract 373; Main 4817; Figures 4; Tables 0; Refs 36; Supplemental tables 4

Proteomics data associated with this manuscript have been uploaded to the ProteomeXChange consortium under accession PXD006353 and is available for review with the following login details:

User: reviewer64884@ebi.ac.uk

Password: 5WwmCUR8

Word Counts Abstract 197, main 4959, figures 4, tables 0, references 32.

Keywords: cerebrospinal fluid, exosome, extracellular vesicle, extraction, origin

ABSTRACT

Cerebrospinal fluid (CSF) extracellular vesicles (EVs) show promise as a source of neurological disease biomarkers, though their precise origin is poorly understood. Current extraction techniques produce disappointing yield and purity. This study describes the application of ultrafiltration liquid chromatography (UFLC) to CSF-EVs, compared with ultracentrifugation (UC), and explores CSF-EV origin. EVs were extracted from human CSF by UC and UFLC and characterized using nanoparticle tracking analysis, electron microscopy and immunoblotting. EV and CSF proteomes were analysed by LC-MS/MS. UFLC-isolated particles have size, morphology and marker expression characteristic of EVs. UFLC provides greater EV yield (UFLC $7.90 \times 10^8 \pm \text{SD } 1.31 \times 10^8$ EVs/mL CSF, UC $1.06 \times 10^8 \pm 0.57 \times 10^8$ $p < 0.001$). UFLC enhances purity, proteomic depth (UFLC 704 ± 52 , UC 340 ± 57 identifications, $p < 0.01$) and consistency of quantification (CV 17% vs 23%). EVs contain more intracellular proteins (OR 2.63 $p < 0.001$) and fewer plasma proteins than CSF (OR 0.60, $p < 0.001$). CSF and EV-enriched proteomes show overrepresentation of brain-specific proteins (EV OR 3.18, $p < 0.001$; CSF OR 3.37, $p < 0.001$). Overrepresentation of cerebral white matter (OR 1.99, $p = 0.015$) and choroid plexus proteins (OR 1.87, $p < 0.001$) is observed in EVs. UFLC provides improves yield and purity of CSF-EVs. The EV-enriched proteome better reflects the intracellular and white matter proteome than whole CSF.

Statement of significance

This study used LC-MS/MS to characterise the proteome of CSF extracellular vesicles extracted using ultrafiltration liquid chromatography size-based extraction method with advantages over ultracentrifugation. A statistical approach was employed to integrate open source databases of brain structure and organ tissue gene expression and proteomic datasets, to infer the provenance of CSF extracellular vesicles. This approach suggested a predominance of choroid plexus and white matter origin for CSF extracellular vesicles and also demonstrated the preponderance of intracellular and non-serum proteins in CSF extracellular vesicles compared with whole CSF. These data support the relevance of extracellular vesicles as a source of biomarkers for CNS diseases, particularly those in which white matter pathology is a feature.

Introduction

Extracellular vesicles (EVs) are membrane-bound vesicles measuring 50-1000nm, that carry a cargo of protein and nucleic acid, formed in multivesicular bodies (exosomes) or by outward budding of the plasma membrane (microvesicles).^[1] Almost all cell types release EVs, with broad-ranging functions including intercellular communication and disposal of unwanted proteins.^[1-3] Neurons, glia and choroid plexus epithelium of the CNS secrete EVs; some reach the cerebrospinal fluid (CSF).^[4-7]

In vitro evidence supports a role for EVs in the pathogenesis of neurological diseases, particularly neurodegeneration, primarily as mediators of intercellular spread of misfolded protein species hypothesized to underlie the observed spatial progression of clinical and histopathological findings (though evidence suggests an ameliorative role in some neurodegenerative diseases).^[8]

CSF-derived EVs contain neuronal, oligodendroglial, astrocytic and microglial proteins, as well as proteins implicated in neurodegeneration,^[7] but less is known about their origin. A major challenge in the study of CSF-EVs is their relative paucity, numbering fewer than 10×10^9 per mL in adults.^[9] Standard methods of EV extraction using ultracentrifugation (UC) were developed for high-volume applications such as extraction from cell culture medium and have poor EV yield and high inter- and intra-operator variability.^[10-11] UC remains the only method capable of

separating exosomes from other EV types, using differential centrifugation or density gradients, and has been employed in previous studies of CSF EVs.^[7, 12-16]

Alternative approaches employing precipitation, though producing higher yield, preclude downstream analytical methods such as proteomic analysis due to the use of PEG and co-precipitation of large quantities of contaminating protein.^[10] Size exclusion methods have provided high yield from cell culture medium and plasma,^[10, 17] but have not been systematically examined in CSF.

To elucidate tissue and CNS structural origin of the CSF and EV-enriched proteome, this study compared a size exclusion-based technique (ultrafiltration-liquid chromatography, UFLC),^[17] with a standardised UC method, in terms of yield, purity and proteomic coverage using liquid chromatography tandem mass spectrometry (LC-MS/MS).

Materials and Methods

Patient samples

CSF samples were obtained from individuals undergoing lumbar puncture as part of routine care on the neurosciences ward. Approval was obtained from Ethics Committee South Central – Oxford C (reference 09/H0606/5+5) and NRES Committee South Central – Hampshire B (14/SC/1248). All participants provided informed consent. Diagnoses included multiple sclerosis, idiopathic intracranial hypertension, narcolepsy, autoimmune encephalopathies,

normal pressure hydrocephalus and peripheral neuropathy. In accordance with consensus guidelines for biomarker development,^[18] CSF was sampled directly into polypropylene containers, centrifuged at 4°C at 3000rpm for 10 minutes and the supernatant aliquoted and stored at -80°C within 1 hour of sampling until use. Samples with visible bloodstaining or red cell count >200 cells/mm³ were excluded.

EV extraction

Samples were thawed on ice and combined. Pooled samples were divided into 8mL aliquots. All analysis comparing UC, UFLC and CSF was performed on paired pools in order to enable direct comparisons. Pools underwent centrifugation at 1200xg for 10 minutes to remove cell debris or large aggregates then filtered through a 0.22µm Millex® 33mm polyethersulfone syringe-driven filter (Merck Millipore) to remove larger microvesicles and apoptotic bodies prior to ultrafiltration or ultracentrifugation.

Ultracentrifugation

UC was performed as described previously, with modifications.^[11] Following 0.22µm filtration, samples were centrifuged in an OptimaMax ultracentrifuge (Beckman Coulter; MLS-50 rotor, k-factor=71) at 120,000xg for 120 minutes. The EV pellet was resuspended in 4mL PBS and centrifuged at 120,000xg for 120 minutes. The pellet was then resuspended in 100µL PBS.

Ultrafiltration liquid chromatography

Following 0.22µm filtration, samples were filtered using Amicon® Ultra-15 100kDa molecular weight cut-off (MWCO) centrifugal filters (Merck Millipore) at 3500xg for 8 minutes, washed with 4mLs PBS and centrifuged at 3500xg for 4 minutes. Retentate volume was adjusted to 800µL with PBS, injected into a 24mL size exclusion column packed with sepharose 4 fastflow (mean particle size 90µm, exclusion limit 3×10^7) and eluted with 40mLs PBS at 0.5mL/min using an ÄKTA pure chromatography system (GE Life Sciences). 2mL fractions were collected from 6-40mLs elution volume. EV-containing fractions (2-3) were concentrated for further analysis using Amicon® 10kDa MWCO 4mL centrifugal filters.

EV characterisation

EV size distribution and concentration was ascertained using nanoparticle tracking analysis (NTA) using a NanoSight NS500 (Malvern Ltd) and NTA 2.3 software. Where necessary, samples were diluted in PBS to a concentration of 2×10^8 - 2×10^9 particles/mL. Size distribution was averaged across three 30-60 second recordings per sample.

Immunoblotting

Immunoblotting was performed using the XCell Surelock Mini-Cell. 19.5µL EV samples were mixed with 7.5µL NuPAGE LDS Sample Buffer (4X, Invitrogen) and 3µL NuPAGE Reducing Agent (10X), heated to 95°C for 8 minutes, loaded into 10-well 1.0mm NuPAGE Novex Bis-Tris 4-12% gels and run at 125V for 45-60 minutes in MES SDS running buffer until the dye front reached the bottom of the gel. Proteins were transferred to polyvinylidene fluoride membrane using the iBlot2 gel transfer device (ThermoFisher Scientific) for 7 minutes. Membranes were

blocked for 60 minutes with gentle shaking at room temperature using 5% fat-free milk in TBS with 0.1% Tween (TBS-T).

After blocking, membranes were incubated with primary antibody, anti-syntenin-1 (SDCBP), anti-CD9 or anti-PDCD6IP (Programmed cell death 6-interacting protein; Abcam) at 1:500 dilution overnight at 4°C with gentle shaking. Membranes were washed 3 times in TBS-T for 10 minutes each with gentle shaking prior to addition of secondary antibody, peroxidase-labelled anti-mouse (PDCD6IP) or anti-rabbit (CD9, SDCBP) at 1:5000 dilution for 60 minutes at room temperature. Membranes were washed 3 times in TBS-T for 10 minutes, then incubated with SuperSignal West Pico peroxidase chemiluminescent reagent for 5 minutes at room temperature before development using photographic film, using exposure times of up to 10 minutes.

Transmission Electron Microscopy

10µL of EV solution was applied to freshly glow-discharged carbon-coated 200 mesh copper grids for 2 minutes, blotted with filter paper and stained with 2% uranyl acetate for 10 seconds, blotted and air dried. Grids were imaged in a FEI Tecnai 12 transmission electron microscope at 120kV using a Gatan OneView CMOS camera. Particle size was measured using ImageJ v1.50i (National Institutes of Health, USA).

Sample preparation for proteomic analysis

Samples of whole CSF, concentrated UFLC and UC samples were reduced in 5mM dithiothreitol for 30 minutes at room temperature followed by alkylation with 20mM iodoacetamide for 30 minutes at room temperature and precipitated using chloroform-methanol precipitation. Precipitated protein was resuspended in 50mM triethylammonium bicarbonate with vortexing and sonication for 2 minutes. Samples were digested overnight at 37°C with 300rpm shaking using 400ng of trypsin for UFLC and UC samples and a 1:50 trypsin:protein ratio for whole CSF. Peptide digests were acidified with 1% formic acid, desalted using Sep-Pak C18 cartridges (Waters) and dried by vacuum centrifugation. Peptides from UFLC and UC were resuspended in 10µL buffer A (2% acetonitrile, 0.1% formic acid in water); whole CSF samples were resuspended to a final concentration of 500ng/µL and kept at -20 °C until analysis.

Liquid chromatography tandem mass spectrometry

Peptides were analysed by nano ultra-high-performance liquid chromatography tandem mass spectrometry (nano-UPLC-MS/MS) using a Dionex Ultimate 3000 nanoUPLC, (Thermo Scientific) coupled to an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Scientific). 5µL UFLC, UC or whole CSF peptide digest were loaded onto an EASY-Spray column (75µm×500mm, 2µm particle size, Thermo Scientific) and eluted using a 60-minute gradient starting with 2% acetonitrile, 0.1% formic acid and 5% DMSO, increasing to 35% acetonitrile, 0.1% formic acid and 5% DMSO at a flow rate of 250nL/min. Data were acquired in data dependent mode with a resolution of 120,000 full-width half maximum at m/z 200 in the survey scan (375-1500 m/z) and with EASY-IC using the reagent ion source (202 m/z) for internal calibration. MS/MS spectra

were acquired after precursor isolation in the quadrupole with an isolation window of 1.2Th, dynamic precursor selection (top speed mode) with a 3-second fixed duty cycle time and 60-second dynamic precursor exclusion. Isolated precursor ions were fragmented by CID with a normalised Collision Energy of 35%. Parallelization was enabled and MS/MS spectra were acquired in the linear ion trap for up to 250ms with an ion target of 4000 in rapid scan mode.

Raw MS data were analysed using Progenesis QI for Proteomics v3.0 (Nonlinear Dynamics). MS/MS spectra were searched against the UniProt Homo Sapiens Reference proteome (retrieved 15/11/2016) using Mascot v2.5.1 (Matrix Science) with precursor mass tolerance of 10ppm and a fragment ion tolerance of 0.5Da. Carbamidomethylation of Cysteines was defined as a fixed modification, and deamidation of Asparagine and Glutamine and oxidation of Methionine as variable modifications. Peptides scoring ≥ 20 and FDR $<1\%$ were imported into Progenesis QIP.

Proteins with ≥ 2 unique peptides were included in the comparison of UFLC and UC proteomes. Protein abundance was normalized by Log_2 -transformation followed by centering on the median abundance of the 90% of proteins with the lowest variance across all runs and scaling by median absolute deviation. Missing values were imputed by random sampling from a normal distribution downshifted 1.8 SDs from the mean intensity of normalised values, with SD 0.25 times the SD of normalised values ^[19].

Statistical Analysis

Analysis of relative abundance data was performed in R using Welch's T-test with FDR correction using the Benjamini-Hochberg procedure.^[20] A p-value < 0.05 (or FDR <5%) was taken to denote statistical significance. GO enrichment analysis was performed using TopGO. Comparisons between UFLC and UC used proteins increased in UFLC or UC samples with FDR-adjusted $p < 0.05$ as the foreground list and all proteins identified as the background list.

Matrix effects due to the differing composition of whole CSF and purified EVs are likely to have a major effect on peptide ionization and quantification. Comparisons of the CSF and EV proteomes were performed using the list of proteins identified by Mascot with ≥ 1 unique peptide in ≥ 2 whole CSF or UFLC-extracted EV samples. Proteins identified in whole CSF samples were excluded from the EV list.

GO enrichment analysis comparing EV-enriched and CSF proteomes was performed using a target-background approach at the gene level, with proteins identified by Mascot in only EVs or whole CSF as the foreground list. The background list comprised all proteins identified in CSF and EV samples combined.

Analysis of predicted protein secretion by classical and non-classical secretory pathways was performed using SecretomeP and SignalP,^[21] with a SecretomeP score of > 0.6 taken to indicate likely non-classical secretion, where an N-terminal signal peptide was not predicted by SignalP.

Tissue enrichment analysis was performed using data from the Human Protein Atlas, which compares RNA expression across 33 different organs.^[22] Overrepresentation of cerebral cortex- and liver-enriched proteins (mRNA >5-fold higher than in the next highest tissue) was sought in EV-only or CSF proteins by hypergeometric test, using all tissue-enriched proteins as the background list.

Brain structure enrichment analysis was performed using data from the Allen Brain Atlas microarray survey of the human brain,^[23] which includes normalised data from 3702 samples of 6 brains of differing ages and genders using over 62,000 probes with at least two probes for 93% of genes with EntrezGene IDs. Details of donor brains, structure ontology and normalisation strategy are available in ref.^[23] and via the Allen Brain Atlas data portal (<http://www.brain-map.org>). Tissue samples were grouped according to the Allen Brain Atlas ontology to cerebral cortex, subcortical grey (diencephalon, mesencephalon, metencephalon, myelencephalon) and white matter (telencephalic white matter, mesencephalic white matter, metencephalic white matter and myelencephalic white matter).

Gene expression was compared between structures using the probe showing the maximum variance across all samples for each gene. Proteins showing higher mean expression in each structure (fold change >1.2) were then compared with the EV-only or CSF-only protein list, abstracted to genes, using proteins identified in the whole CSF and CSF-EV proteomic analysis as background. Overrepresentation was analysed using a hypergeometric test. Of the selected structures, only cerebral cortex, diencephalon, mesencephalon, metencephalon,

myelencephalon, telencephalic white matter and choroid plexus had differentially expressed genes and were therefore used in the overrepresentation analysis.

Overrepresentation of EV and CSF proteins in myelin and cortical neurons was performed, using proteomic datasets obtained in previously published studies.^[24-25] Overrepresentation was sought by comparing the whole CSF and EV-enriched proteomes with the proteome identified using mass spectrometry in myelin and laser capture microdissected cortical neurons,^[24-25] using a hypergeometric test comparing with a background of all proteins identified in the whole CSF and CSF-EV proteomic analysis.

Results

Characterisation of UFLC and UC performance

EV isolation and yield

NTA identified the peak particle concentration at 8-12mL elution volume, corresponding to fractions 2-3 and the terminal portion of the column void volume (**Figure 1A**). Most protein eluted at 16-30mL volume, corresponding to fractions 5-11 (**Figure 1A**). NTA of concentrated EV peak fractions demonstrated particles with size distribution characteristic of a mixed population of EVs (**Figure 1B**). Immunoblotting demonstrated characteristic EV markers CD9 antigen, syntenin-1 and programmed cell death 6-interacting protein (**Figure 1C**). Transmission electron microscopy (TEM) of concentrated fractions 2-3 demonstrated particles with morphology consistent with EVs (**Figure 1D**). Notably EVs of diameter > 220nm would be excluded by 0.22µm filtration prior to UC or UFLC.

The total number of EVs recovered from 8mL CSF was >7-fold greater using UFLC than UC, (mean EVs recovered $6.32 \times 10^9 \pm \text{SD } 1.05 \times 10^9$ vs $8.49 \times 10^8 \pm 4.57 \times 10^8$, $p < 0.001$, $n = 4$; **Figure 2A-B**) with comparable size distribution as measured by NTA (modal diameter UC mean = $96.25 \pm \text{SD } 19.69 \text{ nm}$, UFLC $76.25 \pm 8.02 \text{ nm}$, $p = 0.191$; combined median diameter UC = 111.88 nm IQR $80.60\text{-}152.75 \text{ nm}$, UFLC median 90.90 nm , IQR $67.73\text{-}123.69 \text{ nm}$; **Figure 2A**). Particle sizes as measured by TEM were also similar, although smaller than estimates obtained by NTA (UC median diameter 49.33 nm , IQR $39.89\text{-}101.03 \text{ nm}$; UFLC median 49.26 , IQR $39.71\text{-}62.31 \text{ nm}$).

The consistency of EV yield was improved in UFLC compared to UC, with CV 16.7% and 53.8% respectively. Characteristic EV markers including CD9 antigen, programmed cell death 6 interacting protein, syntaxin-1, CD81 antigen, flotillin-1 and tumour susceptibility gene-101 were identified in analysis of LC-MS/MS data consistently in both UFLC and UC isolated EVs (**supplementary table 1**). In keeping with the increased number of EVs demonstrated by NTA, relative abundance of all EV markers was higher in UFLC samples prior to normalisation ($p < 0.01$ for all markers).

Exclusion of contaminating protein

Protein concentration was too low to quantify with conventional methods in UC and UFLC samples. Ratios of the 10 most abundant CSF proteins to the most abundant EV marker CD9 were higher in UC for Albumin, Immunoglobulin Kappa constant, Cystatin C, Transthyretin, Alpha-1-antitrypsin and Transferrin, suggesting improved exclusion of abundant contaminating

proteins using UFLC (n=3, $p < 0.05$, **Figure 2C**). Ratios of lipoprotein markers Apolipoprotein A1 and Apolipoprotein E were higher in UFLC, implying that the size-based extraction method is less effective at excluding lipoprotein particles (**Figure 2C**). These results must be viewed with some caution, since some cell types have been reported to release Apolipoprotein E in association with exosomes,^[26] and other high-abundance CSF proteins might be released in EVs.

To extend this analysis, CD9 ratios of all proteins identified in whole CSF and EV samples (both UC and UFLC) were calculated (285 proteins). Overall, mean ratios were higher in UC than UFLC (UC mean $-1.44 \pm \text{SD } 1.32$, UFLC -1.75 ± 1.24 , $p = 0.004$), again implying improved exclusion of abundant proteins by UFLC (**Figure 2D**).

UFLC and UC proteomes

A total of 1083 protein groups were identified with at least two unique peptides in UFLC and UC samples. Label-free quantification of 668 proteins was possible across UFLC and UC samples (**Figure 2D**). The CV of label-free protein quantification in UFLC was lower than UC (median CV UFLC 16%, UC 21%). Significantly more Mascot identifications per run were achieved in UFLC samples than UC (704 ± 52 vs 340 ± 57 identifications, $p < 0.01$; **Figure 2E**). Following normalisation, 142 proteins were differentially abundant with FDR-adjusted $p < 0.05$ (**Figure 2F** and **supplementary table 2**). GO enrichment analysis of differentially abundant proteins compared with a background of all proteins identified in UC and UFLC samples indicated overrepresentation of terms (FDR-adjusted $p < 0.05$) “basal lamina” and “extracellular matrix

structural component” in UFLC, and “intraciliary transport particle B” and “cilium morphogenesis” in UC.

Proteins that have been previously found to be enriched in EV populations isolated at different centrifugation speeds were examined for differential abundance in UC and UFLC samples.^[12] There were no significant differences at the FDR-adjusted $p < 0.05$ threshold, though actinin-4 (ACTN4), which is enriched in medium and large EVs pelleting at low speeds (2-10,000g), had higher mean abundance in UC samples, whilst syntenin-1, which is enriched in small EVs pelleting at higher speeds (100,000g), had higher mean abundance in UFLC samples (**Figure 2F**).

The EV-enriched proteome

Stringent criteria were applied for proteins in the EV list, with exclusion of proteins only identified in one sample and those detected also in CSF. The remaining 359 proteins were considered to constitute the EV-enriched proteome for subsequent analysis. All proteins identified in ≥ 2 whole CSF (including those also found in EVs), were considered to constitute the CSF proteome. All proteins identified in UFLC-EV and CSF samples are given in **Supplementary Table S-3**.

Comparison of the EV-enriched proteome with databases of previously identified EV proteins, Exocarta and Vesiclepedia, demonstrated a high degree of overlap, with only 28 proteins identified in the EV-enriched proteome that were not included in Vesiclepedia or Exocarta

(**Figure 3A**). Of the 100 most frequently identified EV proteins, including EV marker proteins, 45 were included in the EV-enriched proteome, compared with 22 identified in whole CSF (OR 3.32, $p < 0.001$; **Figure 3B**).

GO enrichment analysis was performed, comparing EV-enriched proteome, CSF proteome and overlapping proteomes against the background of all proteins identified. Cellular component terms identified differences in the origin of CSF and EV proteins, with the EV-enriched proteome enriched for extracellular vesicle, plasma membrane and cytoplasmic terms. CSF-enriched terms related to the extracellular space, complement and coagulation machinery (**Figure 4A-C**).

Biological process and molecular function terms in EVs related to extracellular vesicles, GTPase activity, and cell-cell adhesion, whilst those in CSF related to endopeptidases, extracellular matrix binding, complement, and humoral immunity. Notably, there were some immune system-related terms in the EV-enriched proteome (“MHC class II complex binding”) and some intracellular terms in the whole CSF proteome (“cytoplasmic membrane-bound organelle lumen”). Interestingly, though not meeting the 5% FDR threshold, “myelin sheath” was enriched in the EV-enriched proteome, whilst “neuron recognition” and “synapse membrane” were enriched in the whole CSF proteome.

GO enrichment analysis of the overlapping proteome identified terms that also found in, or related to, whole CSF and EV-enriched GO terms. Enriched GO terms with FDR-adjusted $p < 0.05$

are given in **Figure 4** and top 20 cellular component, biological process and molecular function terms are given in **Supplementary Table S-4**.

In agreement with GO enrichment analysis, the proportion of proteins with predicted intracellular origin (according to the Human Protein Atlas) was higher in the EV-enriched proteome compared to whole CSF (OR 2.63, $p < 0.001$) or the overlapping proteome (OR 2.49, $p < 0.001$; **Figure 5A**). In support of these findings, the proportion of predicted secretory proteins using SecretomeP and SignalP was higher in the CSF proteome compared with the EV-enriched proteome (OR 0.67, $p < 0.001$; **Figure 5C**); of proteins predicted to be secreted, those predicted to be non-classically secreted was much higher in the EV-enriched proteome compared with the CSF proteome (OR 8.19, $p < 0.001$; **Figure 5C**).

EV proteins were also less likely to be found in the plasma proteome compared to either whole CSF (OR 0.60, $p < 0.001$) or the overlapping proteome (OR 0.32, $p < 0.001$; **Figure 5D**), using data from the Human Protein Atlas.

Considering markers of specific organelles, the EV-enriched proteome contained the early endosome marker RAB5C (as well as eight other RAB proteins), multivesicular body markers CD81 and CD82 and the plasma membrane marker ATP1A1 but did not contain any of the endoplasmic reticulum markers calreticulin, calnexin and ER lumen protein-retaining receptor 1 (KDEL), or lysosomal proteins LAMP1, LAMP2A and LAMP2B.

Analysis of additional Human Protein Atlas data (comparing RNA expression in different tissues to identify a tissue-enriched proteome) demonstrated overrepresentation of cerebral cortex-enriched proteins in both EV (OR 3.18, $p < 0.001$) and CSF proteomes (OR 3.37, $p < 0.001$; **Figure 5E**). Twenty-two cerebral cortex-enriched proteins were identified in EV samples, 33 in whole CSF and 10 in the overlapping proteome. Overrepresentation analysis for liver-enriched proteins (a major contributor to the blood proteome through production of albumin, clotting factors and complement),^[22] demonstrated enrichment of liver proteins in the overlapping proteome (OR 44.83, $p < 0.001$), whole CSF (OR 19.47, $p < 0.001$) and, to a lesser extent, EV-enriched proteomes (OR 2.58, $p = 0.041$; **Figure 5E**).

To determine the relative contribution of different brain regions to the EV and CSF proteomes, microarray data from the Allen Brain Atlas were used. Overrepresentation analysis of proteins with higher expression in brain structures was compared in the EV-enriched, overlapping and whole CSF datasets, using all proteins identified in the proteomic analysis as a background list.

This analysis indicated overrepresentation of genes with higher expression in telencephalic white matter in the EV-enriched proteome compared to other brain regions (OR 1.99, $p = 0.015$; **Figure 5F**). The overlapping genes predominantly comprised those expressed by oligodendrocytes or involved in myelination (including carbonic anhydrase 2 (CA2), CD9, Dynein heavy chain domain-containing protein 1 (DNHD1), Ectonucleotide pyrophosphatase/phosphodiesterase family member 6 (ENPP6), Band 4.1-like protein 2

(EPB41L2), Hyaluronan and proteoglycan link protein 2 (HAPLN2), Kallikrein-6 (KLK6), Pleckstrin homology domain-containing family B member 1 (PLEKHB1) and Protein lifeguard 3 (TMBIM1))^[27-33] as well as those expressed on axons (carbonic anhydrase 14 (CA14)),^[27] astrocytes (aquaporin-1 (AQP1)) and microglia (HLA class II histocompatibility antigen, DR alpha chain (HLA-DRA)).^[34]

Cerebral cortex proteins were overrepresented in the whole CSF proteome compared to the EV-enriched proteome (OR 10.38, $p=0.007$). Choroid plexus proteins were also enriched in the EV-enriched proteome (OR 1.87, $p<0.001$; **Figure 5F**). Choroid plexus proteins found in the EV-enriched proteome included several previously shown to be expressed in choroid plexus including an isoform of transthyretin (TTR), fibulin1 (FBLN1) and laminin subunit alpha-5 (LAMA5).^[35-36] Other extracellular matrix proteins such as collagen alpha-2(VI) chain (COL6A2) and laminin alpha, beta and gamma subunits as well as nine solute transporters.

There was no significant brain region enrichment identified in the overlapping proteome.

In order to provide additional validation for these findings, additional overrepresentation analysis was performed using a dataset of 673 myelin proteins identified from human *post mortem* frontal lobe tissue.^[25] This demonstrated significant enrichment in the EV-enriched but not whole CSF proteome, compared to the background of all proteins identified in the proteomic analysis, supporting the earlier enrichment analysis (EV OR 2.10, $p=0.001$; CSF OR 0.65, $p=0.977$, **Figure 5G**).

A similar analysis was performed in which overrepresentation of EV-enriched and whole CSF proteins was sought amongst 464 proteins identified in cerebral cortical neurons isolated using laser capture microdissection (LCM)^[24] in order to validate the finding of cerebral cortex enrichment in whole CSF. This did not support the earlier findings, instead demonstrating overrepresentation of cerebral cortex proteins in the EV-enriched proteome (EV OR 1.38, $p=0.006$; CSF OR 0.93, $p=0.761$, **Figure 5H**). Further analysis of an additional dataset comprising 3106 proteins identified in proteomic analysis of LCM-isolated hippocampal neurons from healthy controls and patients with Alzheimer's disease also demonstrated overrepresentation of hippocampal proteins in the EV-enriched proteome (EV OR 1.80, $p<0.001$; CSF OR 0.57, $p=0.999$; **Figure 5I**).

On examination of the dataset for current candidate neurodegenerative disease biomarkers, including amyloid beta A4 protein (APP), tau, S100 proteins, 14-3-3 proteins, neurogranin, presenilin 1 and 2, TAR DNA binding protein-43, alpha synuclein, transforming growth factor beta (TGFB), chitinase proteins (CHI3L1, CHI3L2 and CHIT1), major prion protein (PRNP) and neurofilament proteins,^[37] only amyloid precursor protein, four S100 protein isoforms and four 14-3-3 protein isoforms were found in the EV-enriched list. Chitinase 3-like protein 1 and TGFB-1 were detected in the overlapping list, and PSEN1, chitinase 3-like protein 2, TREM2 and an additional isoform of APP were detected exclusively in whole CSF. In UC samples (including proteins overlapping with CSF), only APP, five 14-3-3 proteins and six S100 proteins were detected

Discussion

UFLC of CSF led to isolation of particles with size distribution and morphology typical of EVs, expressing characteristic EV markers. Extraction of EVs by UFLC led to 7-fold higher recovery compared to UC. UFLC outperformed UC in its exclusion of abundant soluble CSF proteins. As a consequence of improved EV yield, LC-MS/MS of UFLC samples resulted in more protein identifications and improved consistency of quantification compared to UC. No systematic differences in GO terms, or in the levels of markers associated with specific EV subpopulations were identified. UFLC was notably less effective at excluding lipoproteins, suggesting co-elution of lipoproteins with EVs during size exclusion chromatography, though this would not generally be problematic for biomarker development.

The principal limitation of this EV extraction method was the large volume of CSF required to obtain good quality LC-MS/MS data. Adaptations such as ion exchange chromatography or employing smaller size exclusion column volumes may improve this. A number of previous studies, employing different proteomic EV extraction and proteomic methods to characterise the EV proteome have been performed.^[7, 13, 38-39] Two previous studies, one employing UC with MS-based proteomics and the second employing precipitation followed by size-exclusion with aptamer-based proteomics, have used somewhat smaller volumes of (6 and 5 mLs CSF, respectively).^[7, 38] However, these volumes remain well above that typically extracted during

clinical lumbar puncture and, given the improvements in yield and purity, UFLC is the method of choice when compared to UC for biomarker discovery projects involving the characterization of a pooled CSF-EV proteome, enabling more reliable in-depth characterisation of EV cargo.^[7, 13, 39]

CSF-EVs have been examined for candidate biomarkers of neurodegenerative diseases in several target-driven studies, which benefit from smaller sample volumes when compared to shotgun proteomic analysis. Disease-associated phosphoforms of tau have been detected by ELISA in CSF EVs from patients with Alzheimer's disease, extracted using density gradient centrifugation.^[15] Decreased levels of alpha synuclein, the major component of Lewy bodies, the typical inclusions of Parkinson's disease and dementia with Lewy bodies, have been detected by electrochemiluminescent assay in CSF EVs from patients with Parkinson's disease, isolated by UC.^[14] Alterations in CSF EV levels of TAR DNA binding protein-43 (TDP-43), the major aggregate component in amyotrophic lateral sclerosis, have also been sought, though no difference between patients and controls was detected in a single small targeted proteomic study, again employing UC for EV extraction.^[16]

Comparison of CSF-EV and whole CSF proteomes is complicated by differences in precursor ion behaviour in the highly differing composition of digested whole CSF and CSF EVs, negating a simple comparison of abundance measured by intensity or spectral counts. The comparison described here was performed using a method based on non-overlapping identifications in EV samples, though the lack of quantitative analysis across the two proteomes is a limitation; alternative approaches that might minimise these matrix effects would include non-MS-based

proteomic analysis, as employed in a previous study of CSF EVs that used an aptamer-based method,^[38] (though the *a priori* selection of proteins used in the aptamer approach limits the range of detectable proteins) or with the use of isobaric labelling methods.

GO enrichment analysis of the EV-enriched proteome identified intracellular component and process terms; concurring with this, analysis of protein class of EV and CSF proteins (a greater proportion of EV proteins being intracellular compared to CSF or the overlapping proteome) and supporting the notion that CSF-EVs might contain biomarkers more reflective of intracellular events than studying whole CSF.

Comparison of the tissue and CNS structural origin of CSF, EV-enriched and overlapping proteomes demonstrated an overrepresentation of brain-enriched proteins in both EV and whole CSF proteomes, suggesting CNS origin of at least a proportion of CSF-EVs. The substantial overrepresentation of liver-enriched proteins observed in whole CSF and the overlapping proteome is likely to reflect the well-documented overspill of highly abundant serum proteins in CSF, also supported by overrepresentation of plasma proteins in the whole CSF proteome when compared to EVs.^[40]

The EV-enriched proteome showed enrichment of proteins with higher RNA expression in cerebral white matter regions and choroid plexus, whilst in whole CSF enrichment of proteins with higher RNA expression in cortex was observed. Although this analysis should be viewed with caution due to lack of concordance between RNA and protein expression, the additional

analysis of a myelin proteomic dataset provided support for the differential enrichment of myelin (i.e. white matter) proteins within EVs (including proteins such as 2'3'-cyclic-nucleotide-phosphodiesterase (CNP)^[4] previously identified in oligodendrocyte-derived EVs). This finding might indicate that the population of CSF-EVs receives a greater contribution from oligodendrocyte EVs compared to those released by other glial cell types or neurons. It should be noted, however, that samples from diverse conditions were included in the analysis, including, for example, samples from patients with CNS demyelination in whom increased oligodendroglial EV release might occur.

The enrichment of cortical proteins in whole CSF and relative lack in the EV-enriched proteome was not recapitulated when validated using a dataset comprising proteomic analysis of cortical neurons. This may reflect differences between the cellular transcriptome and proteome or be a consequence of the fact that the transcriptomic data used are based on tissue homogenates comprising multiple cell types, whereas the proteomic data use a relatively pure population of cells. Although these data do not support a differential contribution of different brain regions and cell types to the CSF-EV population, they do provide additional evidence that different brain regions contribute to the EV-enriched proteome, specifically cerebral white matter and choroid plexus.

Some important neurodegenerative disease biomarkers were identified in EV samples, though only S100 and 14-3-3 proteins were detected exclusively in the EV samples, and a large proportion of current candidates were not detected in proteomic analysis at all, either due to

low abundance or a lack of suitable peptides; it is noteworthy that several candidate biomarkers not identified in this analysis (specifically TDP-43, tau and alpha-synuclein) have previously been detected in CSF EVs using targeted proteomics and immunoassays.^[14-16]

Supporting information

Mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium via PRIDE (PXD006353 and 10.6019/PXD006353). Supplementary table S-1: raw and normalised protein abundance and differential abundance for UC and UFLC samples. Supplementary table S-2: Differentially abundant UFLC and UC proteins, FDR <0.05. Supplementary table S-3: Protein identifications in CSF and UFLC-EV samples. Supplementary table S-4: Top 20 GO terms in CSF and EV-enriched proteomes.

Acknowledgements

IM receives funding from the EU Horizon 2020 consortium project “B-SMART” and Estonian Research Council project PUT618. RF and BMK are supported by the Kennedy Trust Fund. SEAL is supported by the Swedish Research Council (VR Med and EuroNanoMed) and Swedish Society of Medical Research. MRT is funded by Medical Research Council & Motor Neurone Disease Association Lady Edith Wolfson Senior Clinical Fellowship grant MR/K01014X/1.

We are grateful to Dr Errin Johnson (Sir William Dunn School of Pathology, University of Oxford) for providing TEM images and Dr Justin Hean (Department of Physiology, Anatomy and Genetics, University of Oxford) for advice during optimisation of EV extraction.

Competing Interests

MW and SEAL have filed patent applications in relation to EVs and are founders and shareholders in Evox Therapeutics.

References

- [1] S El Andaloussi; I Mager; XO Breakefield; MJ Wood, *Nat Rev Drug Discov* **2013**, *12*, 347-57.
- [2] BT Pan; K Teng; C Wu; M Adam; RM Johnstone, *J Cell Biol* **1985**, *101*, 942-8.
- [3] G Raposo; HW Nijman; W Stoorvogel; R Liejendekker; CV Harding; CJ Melief; HJ Geuze, *J Exp Med* **1996**, *183*, 1161-72.
- [4] EM Kramer-Albers; N Bretz; S Tenzer; C Winterstein; W Mobius; H Berger; KA Nave; H Schild, et al., *Proteomics Clin Appl* **2007**, *1*, 1446-61.
- [5] J Faure; G Lachenal; M Court; J Hirrlinger; C Chatellard-Causse; B Blot; J Grange; G Schoehn, et al., *Mol Cell Neurosci* **2006**, *31*, 642-8.
- [6] M Grapp; A Wrede; M Schweizer; S Huwel; HJ Galla; N Snaidero; M Simons; J Buckers, et al., *Nat Commun* **2013**, *4*, 2123.
- [7] D Chiasserini; JR van Weering; SR Piersma; TV Pham; A Malekzadeh; CE Teunissen; H de Wit; CR Jimenez, *J Proteomics* **2014**, *106*, 191-204.
- [8] AG Thompson; E Gray; SM Heman-Ackah; I Mager; K Talbot; SE Andaloussi; MJ Wood; MR Turner, *Nat Rev Neurol* **2016**, *12*, 346-57.
- [9] A Tietje; KN Maron; Y Wei; DM Feliciano, *PLoS One* **2014**, *9*, e113116.
- [10] RJ Lobb; M Becker; SW Wen; CS Wong; AP Wiegman; A Leimgruber; A Moller, *J Extracell Vesicles* **2015**, *4*, 27031.
- [11] C Thery; S Amigorena; G Raposo; A Clayton, *Curr Protoc Cell Biol* **2006**, *Chapter 3*, Unit 3 22.
- [12] J Kowal; G Arras; M Colombo; M Jouve; JP Morath; B Primdal-Bengtson; F Dingli; D Loew, et al., *Proc Natl Acad Sci U S A* **2016**, *113*, E968-77.
- [13] JM Street; PE Barran; CL Mackay; S Weidt; C Balmforth; TS Walsh; RT Chalmers; DJ Webb, et al., *J Transl Med* **2012**, *10*, 5.
- [14] A Stuenkel; M Kunadt; N Kruse; C Bartels; W Moebius; KM Danzer; B Mollenhauer; A Schneider, *Brain* **2016**, *139*, 481-94.
- [15] S Saman; W Kim; M Raya; Y Visnick; S Miro; S Saman; B Jackson; AC McKee, et al., *J Biol Chem* **2012**, *287*, 3842-9.
- [16] E Feneberg; P Steinacker; S Lehnert; A Schneider; P Walther; DR Thal; M Linsenmeier; AC Ludolph, et al., *Amyotroph Lateral Scler Frontotemporal Degener* **2014**, *15*, 351-6.
- [17] JZ Nordin; Y Lee; P Vader; I Mager; HJ Johansson; W Heusermann; OP Wiklander; M Hallbrink, et al., *Nanomedicine* **2015**, *11*, 879-83.
- [18] CE Teunissen; A Petzold; JL Bennett; FS Berven; L Brundin; M Comabella; D Franciotta; JL Frederiksen, et al., *Neurology* **2009**, *73*, 1914-22.
- [19] EC Keilhauer; MY Hein; M Mann, *Mol Cell Proteomics* **2015**, *14*, 120-35.
- [20] R Core Team *R: A language and environment for statistical computing.*, 3.2.3; R Foundation for Statistical Computing: Vienna, Austria, 2015.
- [21] JD Bendtsen; LJ Jensen; N Blom; G Von Heijne; S Brunak, *Protein Eng Des Sel* **2004**, *17*, 349-56.
- [22] M Uhlen; L Fagerberg; BM Hallstrom; C Lindskog; P Oksvold; A Mardinoglu; A Sivertsson; C Kampf, et al., *Science* **2015**, *347*, 1260419.

- [23] MJ Hawrylycz; ES Lein; AL Guillozet-Bongaarts; EH Shen; L Ng; JA Miller; LN van de Lagemaat; KA Smith, et al., *Nature* **2012**, 489, 391-399.
- [24] ES Drummond; S Nayak; B Ueberheide; T Wisniewski, *Sci Rep* **2015**, 5, 15456.
- [25] A Ishii; R Dutta; GM Wark; SI Hwang; DK Han; BD Trapp; SE Pfeiffer; R Bansal, *Proc Natl Acad Sci U S A* **2009**, 106, 14605-10.
- [26] G van Niel; P Bergam; A Di Cicco; I Hurbain; A Lo Cicero; F Dingli; R Palmulli; C Fort, et al., *Cell Rep* **2015**, 13, 43-51.
- [27] S Parkkila; AK Parkkila; H Rajaniemi; GN Shah; JH Grubb; A Waheed; WS Sly, *Proc Natl Acad Sci U S A* **2001**, 98, 1918-23.
- [28] N Terada; K Baracska; M Kinter; S Melrose; PJ Brophy; C Boucheix; C Bjartmar; G Kidd, et al., *Glia* **2002**, 40, 350-9.
- [29] ML Yang; J Shin; CA Kearns; MM Langworthy; H Snell; MB Walker; B Appel, *Dev Dyn* **2015**, 244, 134-45.
- [30] E Kida; S Palminiello; AA Golabek; M Walus; T Wierzba-Bobrowicz; A Rabe; G Albertini; KE Wisniewski, *J Neuropathol Exp Neurol* **2006**, 65, 664-74.
- [31] L Xiao; D Ohayon; IA McKenzie; A Sinclair-Wilson; JL Wright; AD Fudge; B Emery; H Li, et al., *Nat Neurosci* **2016**, 19, 1210-1217.
- [32] JC Dugas; YC Tai; TP Speed; J Ngai; BA Barres, *J Neurosci* **2006**, 26, 10967-83.
- [33] JD Cahoy; B Emery; A Kaushal; LC Foo; JL Zamanian; KS Christopherson; Y Xing; JL Lubischer, et al., *J Neurosci* **2008**, 28, 264-78.
- [34] S Darmanis; SA Sloan; Y Zhang; M Enge; C Caneda; LM Shuer; MG Hayden Gephart; BA Barres, et al., *Proc Natl Acad Sci U S A* **2015**, 112, 7285-90.
- [35] E Thouvenot; M Lafon-Cazal; E Demetere; P Jouin; J Bockaert; P Marin, *Proteomics* **2006**, 6, 5941-52.
- [36] Y Yin; Y Kikkawa; JL Mudd; WC Skarnes; JR Sanes; JH Miner, *Genesis* **2003**, 36, 114-27.
- [37] P Lewczuk; P Riederer; SE O'Bryant; MM Verbeek; B Dubois; PJ Visser; KA Jellinger; S Engelborghs, et al., *World J Biol Psychiatry* **2018**, 19, 244-328.
- [38] JL Welton; S Loveless; T Stone; C von Ruhland; NP Robertson; A Clayton, *J Extracell Vesicles* **2017**, 6, 1369805.
- [39] MG Harrington; AN Fonteh; E Oborina; P Liao; RP Cowan; G McComb; JN Chavez; J Rush, et al., *Cerebrospinal Fluid Res* **2009**, 6, 10.
- [40] A Guldbrandsen; H Vetthe; Y Farag; E Oveland; H Garberg; M Berle; KM Myhr; JA Opsahl, et al., *Mol Cell Proteomics* **2014**, 13, 3152-63.

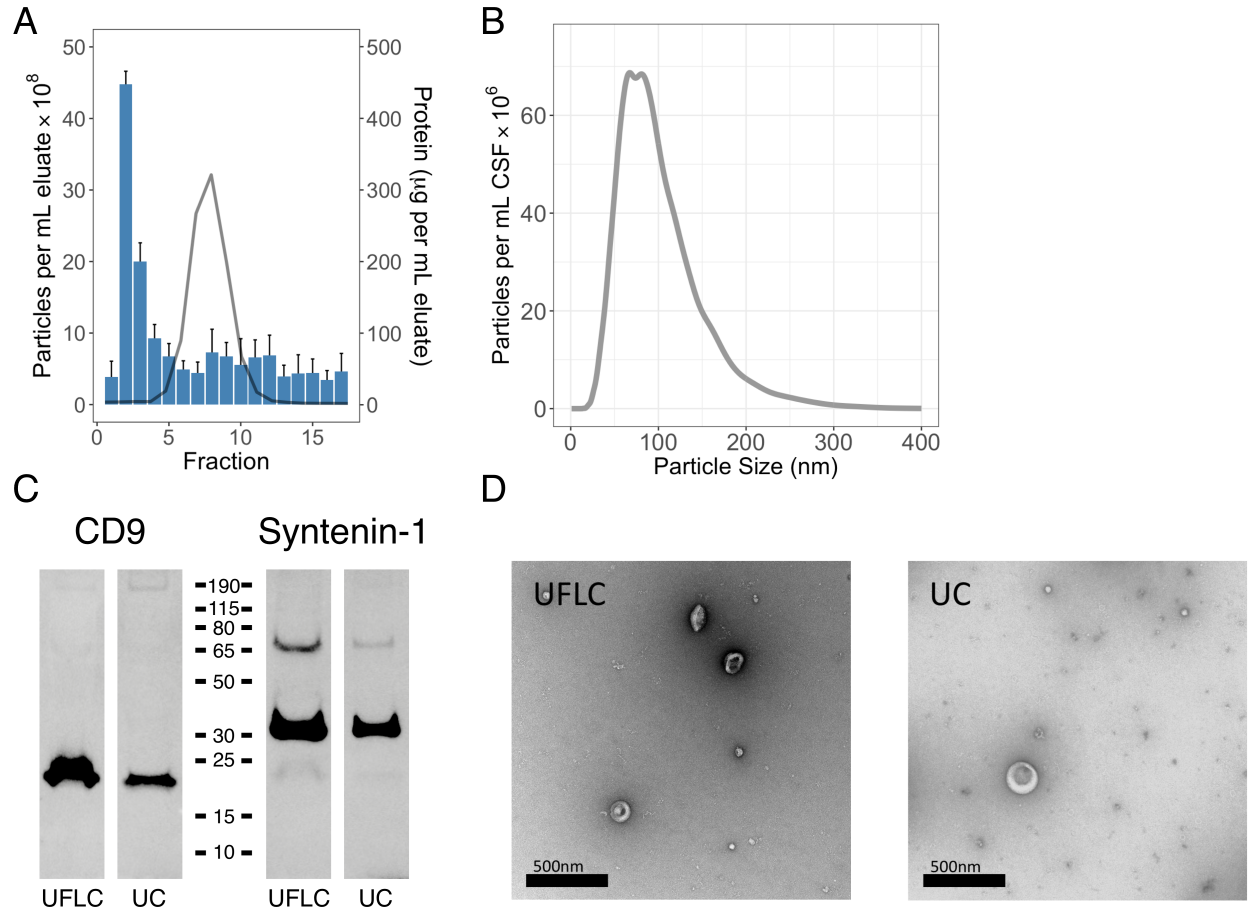


Figure 1 Extraction of EVs by UFLC. **A** Elution profile of UFLC. Blue bars represent NTA data of EV concentration per mL eluate (mean \pm SD, n=4). Grey line indicates mean protein concentration in eluate (mean \pm SD, n=4) **B** Size distribution of extracted EVs following concentration measured by NTA, per mL of CSF starting volume (mean, n=4) **C** Western blot for EV markers programmed cell death-6 interacting protein (PDCD6iP), CD9 antigen and syntenin-1 (SDCBP) in extracted EVs (fractions 2-3) **D** Transmission EM of UFLC- and UC-extracted EV samples.

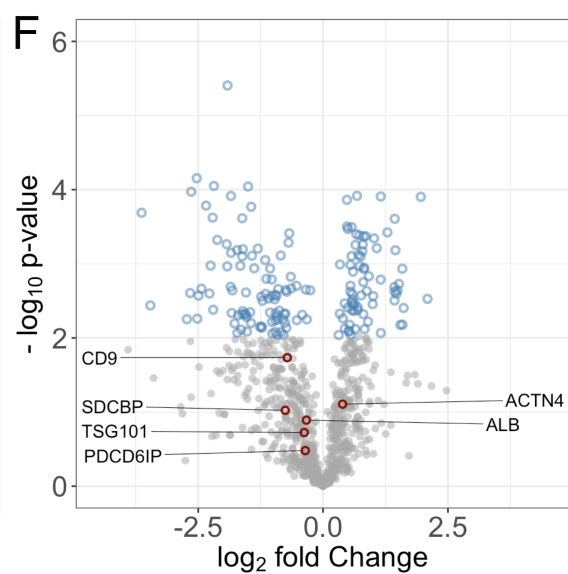
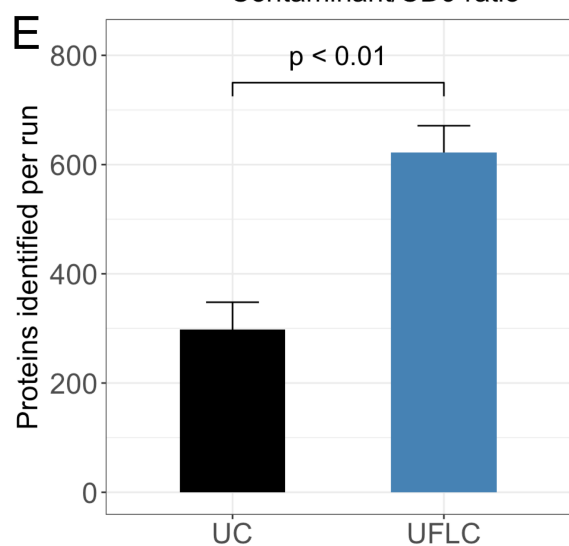
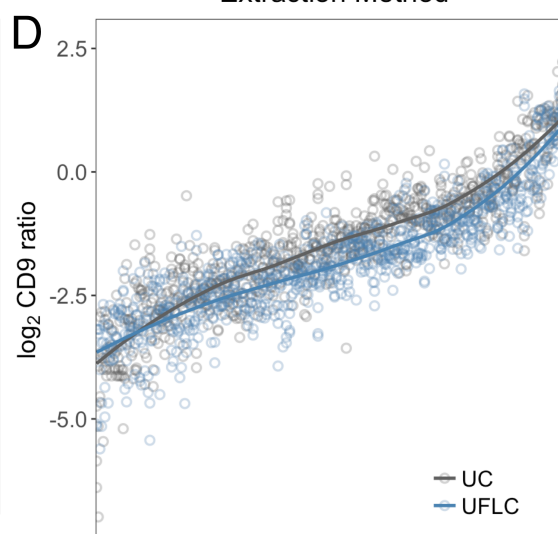
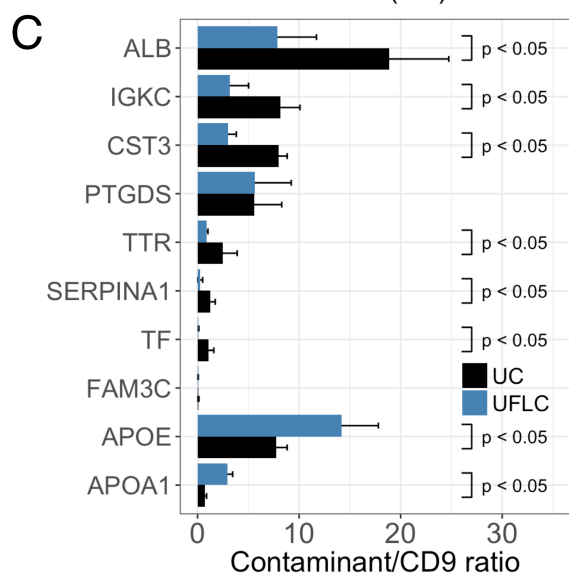
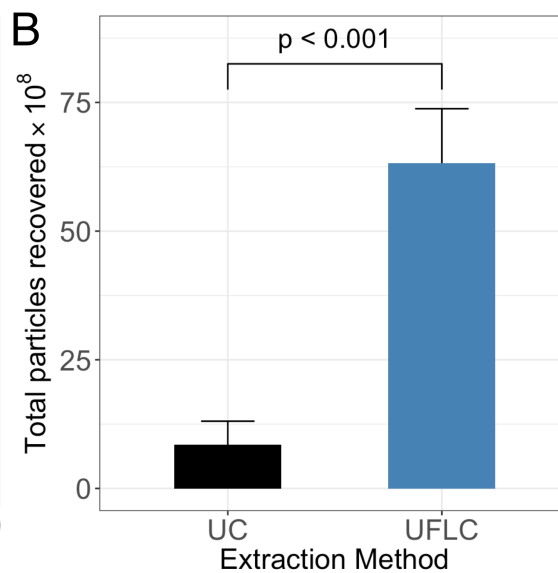
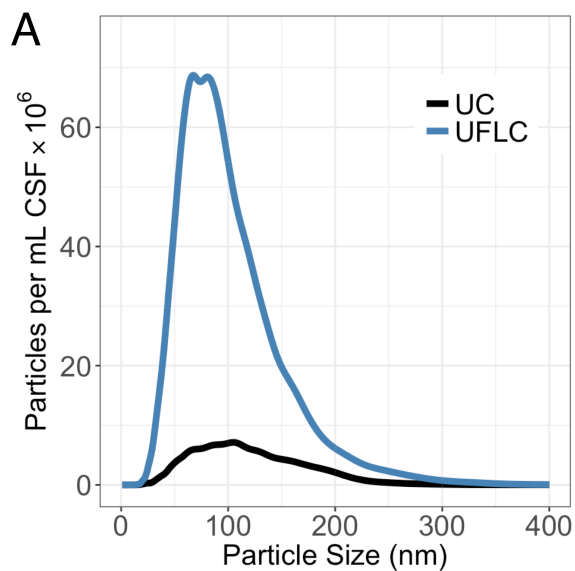


Figure 2 Comparison of UC and UFLC **A** Size distribution of UFLC-extracted EVs, following concentration of fractions 2-3 (blue line) and UC-extracted EVs (black line), adjusted for starting CSF volume, ascertained by NTA (mean, n=4 per method) **B** Total number of EVs recovered by UFLC and UC (mean \pm SD, n=4) **C** Ratio of top 10 most abundant cerebrospinal fluid proteins to CD9 (n=3, mean \pm SD) **D** CD9 ratio for all proteins identified in whole CSF and EV samples. Individual points indicate ratios for each UC or UFLC sample. Line indicates loess fit, proteins ordered by mean CD9 ratio. **E** Number of proteins identified per run of LC-MS/MS (n=3, mean \pm SD) **F** Differences in protein abundance between UFLC and UC-extracted EV samples. Empty blue points indicate proteins with FDR-adjusted $p < 0.05$. EV subtype markers tumour susceptibility gene 101 (TSG101), syntenin-1 (SDCBP) and Albumin (ALB) are indicated with labelled hollow red points.

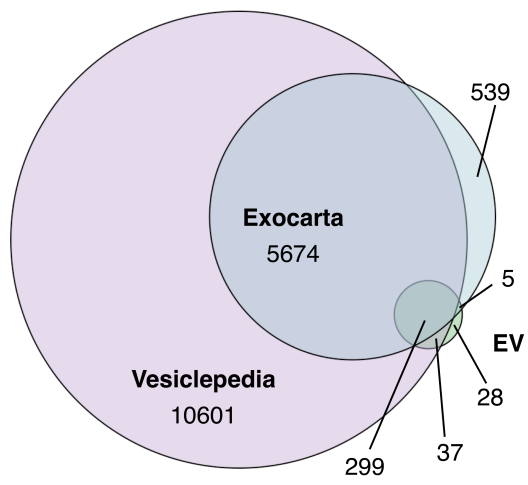
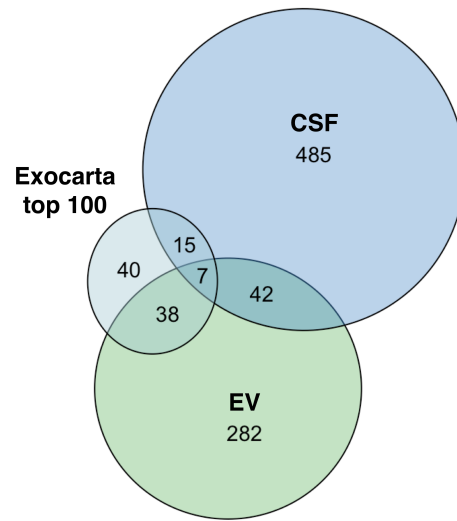
A**B**

Figure 3 A Overlap of the EV-enriched proteome with EV protein databases Exocarta and Vesiclepedia. **B** overlap of CSF and EV proteomes with the top 100 most commonly identified proteins in exosome experiments according to Exocarta.

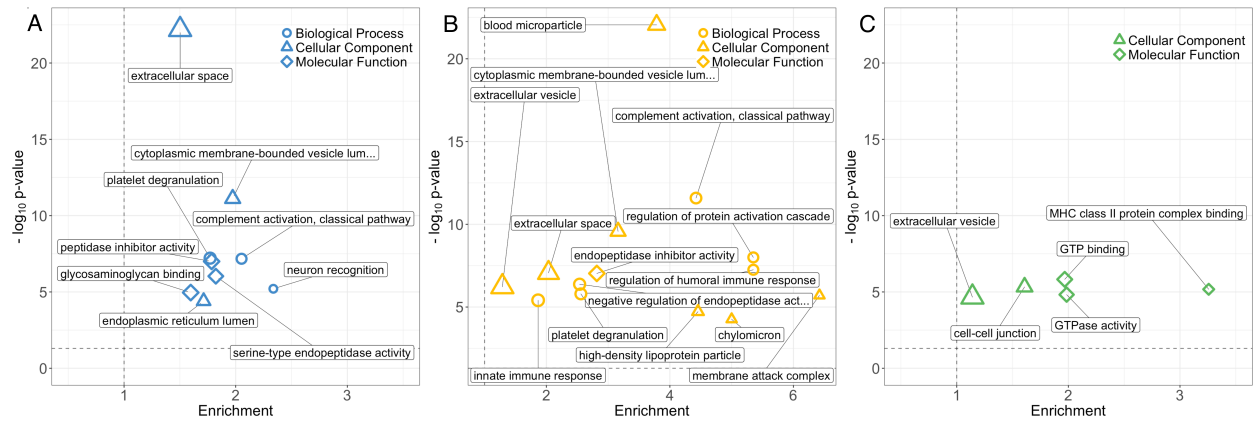


Figure 4 GO enrichment analysis of **A** whole CSF **B** overlapping and **C** EV-enriched proteomes.

Point size proportional to odds ratio. GO terms with FDR-adjusted $p < 0.05$ are shown. Point size proportional to the number of foreground genes annotated to individual terms. Vertical dashed line indicates no enrichment, horizontal line indicates $p = 0.05$.

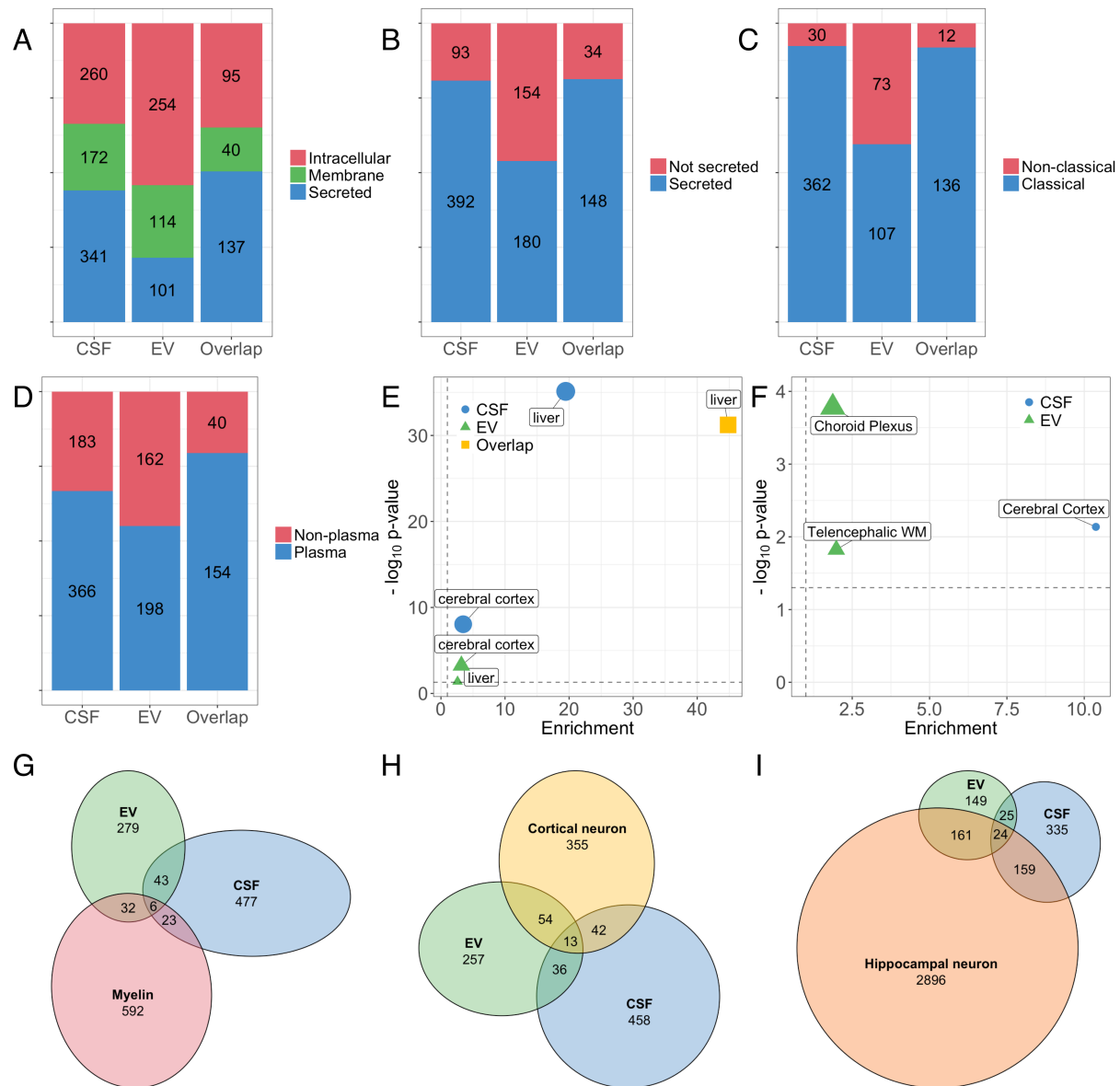


Figure 5 Origin of CSF-EVs. **A** predicted protein class of identified proteins based on Human Protein Atlas data (note that some proteins are annotated to more than one protein class) **B** Predicted secreted and not secreted proteins using SecretomeP and SignalP **C** Predicted classical or non-classical secretion of proteins according to SecretomeP score. **C** Tissue and **D** brain structure enrichment analysis of whole CSF and EV-enriched proteomes. No enrichment was detected in the overlapping proteome. Point size proportional to the number of tissue enriched

proteins per tissue. Vertical dashed line indicates no enrichment, horizontal line indicates $p=0.05$. **E** Proportion of proteins in the EV-enriched, whole CSF and overlapping proteins identified as plasma proteins. **F-H** Overlap of EV enriched and CSF proteomes with proteomes of **F** myelin **G** laser capture microdissected frontal cortical neurons and **H** laser capture microdissected hippocampal neurons.