Leveraging the CSF proteome toward minimally-invasive diagnostics surveillance of brain malignancies

Nicholas Mikolajewicz, Shahbaz Khan, Mara Trifo, Anna Skakdoub, Vladimir Ignatchenko, Sheila Mansouri, Jeffrey Zuccato, Brad E. Zacharia, Michael Glantz, Gelareh Zadeh, Jason Moffat, Thomas Kislinger, and Alireza Mansouri

Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada (N.M., J.M.); Princess Margaret Cancer Centre, Toronto, Ontario, Canada (S.K., V.I., T.K.); Department of Neurosurgery, Penn State Milton S. Hershey Medical Center, Hershey, Pennsylvania, USA (M.T., A.S., B.E.Z., M.G., A.M.); MacFeeters Hamilton Neuro-Oncology Program, Princess Margaret Cancer Centre, University Health Network and University of Toronto, Toronto, Ontario, Canada (S.M., J.Z., G.Z.); Division of Neurosurgery, Department of Surgery, University of Toronto, Toronto, Ontario, Canada (J.Z., G.Z.); Institute for Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada (J.M.); Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada (J.M.); Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada (T.K.)

Corresponding Authors: Alireza Mansouri, MD, MSc, Department of Neurosurgery, Penn State Health, 30 Hope Drive Suite 1200, Hershey, PA, 1701 1, USA (amansouri@pennstatehealth.psu.edu); Thomas Kislinger, PhD, Department of Medical Biophysics, University of Toronto, MaRS Centre, 101 College Street, Room 9-807, Toronto, Ontario, M5G 1L8, Canada (thomas.kislinger@utoronto.ca).

Abstract

Background. Diagnosis and prognostication of intra-axial brain tumors hinges on invasive brain sampling, which carries risk of morbidity. Minimally-invasive sampling of proximal fluids, also known as liquid biopsy, can mitigate this risk. Our objective was to identify diagnostic and prognostic cerebrospinal fluid (CSF) proteomic signatures in glioblastoma (GBM), brain metastases (BM), and primary central nervous system lymphoma (CNSL).

Methods. CSF samples were retrospectively retrieved from the Penn State Neuroscience Biorepository and profiled using shotgun proteomics. Proteomic signatures were identified using machine learning classifiers and survival analyses.

Results. Using 30 µL CSF volumes, we recovered 755 unique proteins across 73 samples. Proteomic-based classifiers identified malignancy with area under the receiver operating characteristic (AUROC) of 0.94 and distinguished between tumor entities with AUROC ≥0.95. More clinically relevant triplex classifiers, comprised of just three proteins, distinguished between tumor entities with AUROC of 0.75–0.89. Novel biomarkers were identified, including GAP43, TFF3 and CACNA2D2, and characterized using single cell RNA sequencing. Survival analyses validated previously implicated prognostic signatures, including blood–brain barrier disruption.

Conclusions. Reliable classification of intra-axial malignancies using low CSF volumes is feasible, allowing for longitudinal tumor surveillance.

Key Points

- High-throughput CSF proteomics on limited volumes from brain tumor patients with intra-ventricular reservoirs is feasible.
- CSF proteomics can be applied toward disease surveillance and may have prognostic utility.
**Importance of the study**

Current approaches to diagnosing brain tumors risk morbidity. The CSF may be an ideal liquid biopsy matrix for mitigating this risk. We report feasibility of high-throughput CSF proteomics on limited volumes from brain tumor patients with intraventricular reservoirs, demonstrate diagnostic and prognostic utility, and explore its applications in practice.

**Background**

Advances in the management of brain malignancies have been limited and patients continue to face a grim prognosis.1–7 Encompassing a broad category that includes high-grade gliomas, brain metastases, and central nervous system lymphomas (CNSL), brain malignancies pose major challenges in clinical management: (1) dependence on invasive tumor tissue sampling for initial histopathology-based diagnosis; (2) imperfect strategies for tumor surveillance following initial therapy; and (3) a lack of clinically actionable, minimally-invasive biomarkers.

At presentation, clinical and imaging parameters alone are not always sufficient for definitively distinguishing high-grade gliomas from brain metastases and CNSL. Given the vastly divergent management for each entity, this necessitates direct tissue acquisition through an invasive neurosurgical procedure which presents with great potential for complication. Even minimally-invasive stereotactic brain tumor sampling has a 4–7% risk of major morbidity.8 This overt dependence on tumor tissue is an even greater challenge during surveillance while on adjuvant therapy and subsequent treatment. Radiation-based treatment regimens can lead to radiation necrosis (RN) in 10–15% of cases while up to 10% of glioblastoma patients can demonstrate pseudo-progression on imaging.9,10 Despite advances in imaging, pathological tissue assessment remains the gold standard approach for distinguishing RN from true tumor progression. Short interval imaging and clinical follow-up is recommended for differentiating pseudo-progression from true tumor progression. Unfortunately, rapid disease progression is not uncommon, at which point many regimens—including enrollment into clinical trials—are no longer feasible due to advanced disease. While longitudinal sampling of the tumor and its microenvironment is necessary for monitoring the expansion of treatment-resistant sub-clones or differentiating tumor recurrence from RN, this is simply not feasible for brain-based pathologies. Thus, it behooves us to develop approaches that help avoid unnecessary surgery while tailoring the specific approach based on tumor prognosis when surgery is necessary.

Commonly referred to as liquid biopsy, sampling of proximal fluids has offered valuable insight in various systemic cancers. In Neuro-Oncology, blood and cerebrospinal fluid (CSF) are the most relevant proximal fluids. While acquisition of blood is associated with a theoretically lower risk of morbidity, the CSF is physiologically the ideal liquid biopsy source for brain tumors, owing to its direct contact with the tumor microenvironment in the central nervous system and limited obstruction by the blood-brain barrier. In routine clinical practice, CSF sampling is central in the management of CNSL and has been used for prognostication of medulloblastoma and germ cell tumors.11 In GBM and brain metastases, although CSF cytology is used clinically to confirm leptomeningeal spread, molecular analyses are currently not part of routine clinical practice for diagnosis, prognostication, and tumor surveillance.

Numerous liquid biopsy-based studies, utilizing a wide variety of molecular assays, have thus far been conducted to develop better diagnostic and prognostic signatures for GBM.11–16 None have been clinically-validated. Circulating tumor DNA (ctDNA), shed predominantly from tumor cell turnover, can be used to genotype GBMs at diagnosis and over the course of the disease. Challenges with ctDNA include its extremely low yield in blood, with only slight improvements detected in the CSF,11 and limited diagnostic alterations distinguishing normal and cancerous ctDNA.17 While providing valuable information, DNA- and RNA-level alterations are unable to predict protein activity, which would be necessary for establishing predictive biomarkers or stratification of patients for use in development of targeted therapeutics.18 Recent proteogenomic analyses of GBM tumor tissue have demonstrated incoherence between mRNA and protein expression, suggesting that the proteome serves as a better representation of disease state and underlying biology.19,20

A comprehensive understanding of GBM and brain metastases through liquid biopsy-based proteomic studies would enable a proactive approach to diagnosis, prognostication, and targeted therapy. This would be paradigm-shifting. As such, application of large-scale proteomic approaches in the realm of liquid biopsy are now imperative. In this study, we report the feasibility of high-throughput proteomics on limited volumes of CSF samples acquired from patients with brain malignancies, describe its diagnostic and prognostic utility, and explore future applications of this approach in routine clinical practice.

**Methods**

The current study adhered to the Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK) guidelines. The completed checklist is provided in Supplemental Appendix 1.

**Software.** Figure preparation: CorelDRAW x8 (Corel); Bioinformatic analyses: R version 4.0.3 (R Foundation for Statistical Computing).
Patient Cohort

Study eligibility.—All patients with a diagnosis of a brain tumor under the care of a physician within the Penn State Hershey Neuroscience Institute who provided informed consent for the Biorepository study, were eligible.

Patient recruitment.—Any patient potentially eligible for the Penn State Hershey Neuroscience Institute Biorepository was first identified by the physician responsible for their care. The physician informed the research coordinator who then independently explained the study to the patient and obtained informed consent. Specimens were only acquired as residuals from samples being collected for routine clinical care or necessary surgical procedures.

Ethics.—The Biorepository study is approved by the Penn State IRB (#2914). Specimens and associated data were released to the protocol “Molecular testing of nervous system cancers as a classification tool” following approval of the Biorepository Data Oversight Committee (#20-0002). Data was released in a de-identified manner via an honest broker system. The patient’s privacy was protected in accordance with both Penn State’s IRB and HIPAA guidelines.

Proteomic Profiling

Sample acquisition.—Most CSF samples were collected from intracranially-implanted CSF reservoirs. The volume withdrawn ranged from 5–15 mL per collection. Samples were aliquoted into 1 mL polypropylene tubes and all samples were processing within 0–4 h after collection, during which the samples were stored at room temperature. The tubes were then spun at 2000 g for 10 min at room temperature to remove cellular debris. Supernatants were maintained in 1 mL aliquots and stored at −80°C. Although higher volumes were available, we only required 30 μL of CSF for reliable proteomics results. In addition, QC samples were created by mixing an equal volume of all 73 samples. Five QC samples were prepared separately, and each QC was run in technical duplicates on the instrument to monitor technical variabilities.

Proteomics.—Protein concentrations were determined by BCA assay (Pierce) and a volume equivalent to 25 μg of protein was used for sample processing, and each sample was spiked in with 2 pmol of yeast invertase (SUC2) as a sample processing control. The samples were denatured and alkylated with DTT and iodoacetamide, respectively. CSF proteins were purified using an adapted MStern technique.21 The samples were bound to a PVDF 96-well MStern plate (Millipore) facilitated by a vacuum suction manifold (Millipore). Adsorbed proteins were washed with 100 mM ammonium bicarbonate (pH = 8) and digested for 4 h at 37°C via the addition of 50 μL of digestion buffer (5% acetonitrile, 100 mM ABC, 1 mM CaCl2) containing 2 μg of trypsin-LysC protease mixture (Promega). The resultant peptides were eluted from the membranes with 50% acetonitrile, lyophilized and desalted with C18 solid-phase extraction tips prepared in-house. 10 μL of purified peptides was spiked with 1 μL of indexed retention time (iRT) (Biognosys) peptide standard. Overall, 11 μL of peptides were loaded onto a 2 cm PepMap Acclaim trap column (Thermo Scientific) using an Easy1000 nanoLC (Thermo Scientific). The peptides were separated and detected against a two-hour reversed-phase gradient using a 50 cm EasySpray analytical C18 column coupled by electrospray ionization to a Q-Exactive HF Orbitrap mass spectrometer (Thermo Scientific) operating in a Top 20 data-dependent acquisition mode. MS1 data was acquired at a resolution of 120,000 with an AGC target of 1e6 ions and a maximum fill time of 40 ms. MS2 data was acquired at a resolution of 30,000 with an AGC target of 2e5 ions and a maximum fill time of 55 ms. A dynamic exclusion of 20 s was enabled, the S-lens RF was set to 59% and the normalized collision energy was set to 27%. The acquired raw data was searched using Maxquant (version 1.6.3.3) against a UniProt complete human protein sequence database (v2020_05) also including yeast invertase (SUC2) and iRT standard peptides.22 Two missed cleavages were permitted along with the fixed carbamidomethyl modification of cysteines, the variable oxidation of methionine and variable acetylation of the protein N-terminus. Relative label-free protein quantitation was calculated using MS1-level peak integration along with the matching-between-runs feature enabling a 2 min retention time matching window. False discovery rate (FDR) was set to 1% for peptide spectral matches and protein identification using a target-decoy strategy. The protein groups file was filtered for proteins detected by a minimum of two peptides and then used to carry out further analysis. Missing LFQ values were imputed with normalized iBAQ intensities.23

Bioinformatic Analyses

Data sources.—Proteomic data from Schmid et al.16 was obtained from ProteomeXchange (ID: PXD021984); Bader et al.24 from ProteomeXchange (ID: PXD016278); Bereman et al.25 from Dataset 1 file of electronic supplementary material of original publication; and Stoop et al.26 from Supporting Information File 2 of original publication. Single cell RNA seq data from Wei et al.27 was obtained from Gene Expression Omnibus (GEO; accession number GSE181304); Neftel et al.28 from GEO (accession number GSE131928); and Kim et al.29 from GEO (accession number GSE131907).

Data preprocessing.—Protein x patient intensity matrices (Supplementary Table S1) were loaded into a Seurat object (Seurat 4.0.4 R package30–33) and sample normalization (column-wise) and protein scaling (row-wise) of log-transformed data was performed. Protein identified in at least 70% of samples within a diagnostic group were retained for downstream analysis.16 No imputations for missing values were performed. For each sample, meta data included a patient identifier, clinical diagnosis (NPH, GBM, BM, CNSL), age, sex, CSF cytology (presence of malignant cells), and leptomeningeal status, alive status...
and survival time. Data from the Schmid cohort were processed identically, however the only meta data that were available were clinical diagnosis, age and sex.\textsuperscript{16}

**Uniform manifold approximation and projection.**—To generate a two-dimensional representation of CSF proteomes, the scaled protein × patient intensity matrix was dimensionally-reduced using robust principal component analysis [\textit{PcaHubert},..., \( k = 50 \), \( k_{\text{max}} = 50 \), \( \text{maxdir} = 100 \), \texttt{signflip = T}), \texttt{rrcov v1.6-0 R package}\textsuperscript{34}] and the top 30 principal components were used for UMAP embedding [\texttt{RunUMAP},...\texttt{, dims = 1:30}, \texttt{Seurat package}].

**Differential protein analysis.**—Differentially-expressed proteins between two groups were identified by two-sided unpaired Wilcoxon test (\textit{wilcoxauc function, presto v1.0.0 R package}\textsuperscript{35}). To ensure results were robust to outliers, we used a resampling procedure that involved 100 iterations of differential expression analysis performed on 95\% randomly sampled subsets of data. Protein that were significant at 5\% false discovery rate (FDR; Benjamini-Hochberg method implemented using \texttt{p.adjust function, stats v4.0.3 R package}) across all iterations were used in downstream analyses (Supplementary Table S2).

**Functional annotation.**—Protein signatures were functionally annotated by performing hypergeometric overrepresentation analysis (\texttt{fora function, fgsea v1.14.0 R package}\textsuperscript{36}) using Gene Ontology (GO) biological processes (BP) and cellular components (CC), protein interaction database (PID), HALLMARK gene sets,\textsuperscript{37} and Reactome\textsuperscript{38} databases. Enrichments were ranked by Benjamini-Hochberg-adjusted \( P \)-values (\( q \)-value), and the top 5 annotations for each database were shown.

**Classification models.**—For LR classification models (Supplementary Table S3; Model 1, Supplementary Table S4; Model 1), training and testing cohorts were generated by randomly splitting CSF samples into 70\% and 30\% subsets, respectively. The training set was used to train a binomial generalized linear model (GLM) using \texttt{bayesglm(..., family = "binomial"; \( \text{maxit} = 500 \)) (\texttt{arm R package, v1.12-2}\textsuperscript{39}) and the AUROC performance was evaluated in the test set (\texttt{performance function, ROCR v1.1-11 R package}).\textsuperscript{40} Given our smaller cohort size, we confirmed that performance was not influenced by choice of partition ratio (50:50, 60:40 and 80:20 train:test splits yielded similar results). This resampling procedure was repeated over 200 iterations to obtain stable estimates of the median AUROC and corresponding 95\% confidence intervals. ROC curves for each iteration were generated using results from \texttt{performance(..., measure = "tpr", x.measure = "fpr"),} and the median ROC curve summarizing overall performance was computed by aggregating the true positive rate (TPR) and false positive rate (FPR) estimates across all iterations. For each model, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) estimates were obtained using cut-offs at which TPR + (1 – FPR) were maximized.

When indicated (Supplementary Figure S3–H), classifiers were also adjusted for leptomeningeal status, presence of malignant cytology in CSF, sex, and age.

For protein signature-based classifiers, (Figure 2, Supplementary Figures S3–S5, Models 1 and 2 in Supplementary Table S3), signatures were aggregated separately for train and test sets using \texttt{gsval(..., method = "gsva") (GSVA R package, v1.36.2}\textsuperscript{41}). Given our limited sample size and goal to also characterize biology associated with malignancy, signature discovery (ie, differential expression analyses, described above) was performed on the total cohort (not training cohort). While in certain contexts such an approach risks leaking information between training and testing cohorts, the re-sampling procedure used in deriving each signature was tailored to minimize cohort-specific biases. Furthermore, in addition to evaluating the performance of each model on the test cohort derived from the current study’s patient samples, we evaluated the external validity of each trained model on matched samples from the entire Schmid et al.\textsuperscript{16} cohort.

For uniplex (single-protein) and triplex (three-protein) classifiers (Figure 3, Supplementary Figure S5), protein intensities were not aggregated prior to training the model, and instead intensity values were used as model inputs. For triplex classifiers, the top performing protein combinations were rank-ordered by the average AUROC across GBM-vs-other, BM-vs-other, and CNSL-vs-other classifiers.

To ensure the robustness of the LR models described above, we independently evaluated L1-regularized (ie, Lasso) LR models (\texttt{glmnet function, glmnet R package v1.36.2}) using the leave-one-out cross-validation procedure implemented in \texttt{caret (care R package v6.0-88)}. For these models, AUROC estimates were calculated and 95\% CI were computed using 2000 bootstrap replicates (\texttt{ci.auc function, pROC R package v1.18.0}) (Supplementary Table S3; Model 2, Supplementary Table S4; Model 2).

**Survival analysis.**—The prognostic value of individual protein was evaluated using univariate Cox proportional hazards regression models (\texttt{coxph function, survival v2.13 R package}). The resulting hazard ratios (HR) were visualized on a volcano plot, with the top hits indicated. Given our limited statistical power for survival analyses, \( P \)-values with no multiple testing corrections were reported to highlight the strongest associations with survival. Survival-associated pathways were determined by performing gene set enrichment analysis (GSEA; \texttt{fgsea function, fgsea v1.14.0 R package}\textsuperscript{36}) on HR-ranked proteins. GSEA plots were generated using the \texttt{plotEnrichment function (fgsea R package)} on HR-ranked proteins. Kaplan Meier survival curves showing patient survival between pathway-stratified groups (high vs. low; split at median) were generated by aggregating protein signature scores with \texttt{gsval(..., method = "gsva") (GSVA R package, v1.36.2)\textsuperscript{41}} and visualizing survival with \texttt{ggsurvplot(...) (survminer R package, v0.4.9).}\textsuperscript{42}

**Single cell transcriptomic analysis.**—ScRNAseq data sets were normalized, scaled, dimensionally-reduced and visualized on a UMAP using the \texttt{Seurat} (version 4.0.4) workflow.\textsuperscript{30–33} In brief, count matrices were loaded into a
Seurat object and normalized using NormalizeData(..., normalization.method = "LogNormalize", scale.factor = 10000). Variable features were identified using FindVariableFeatures(..., selection.method = "mvp", mean.cutoff = c(0,1.8), dispersion.cutoff = c(1,Inf)) and then data were scaled using ScaleData(...). Principal component analysis and UMAP embedding was performed using RunPCA(...) and RunUMAP(..., dims = 1:30), respectively. Metadata from original publications were used to annotate cell types, and gene expression was visualized on a UMAP using FeaturePlot(...).

Data visualization.—Unless otherwise specified, the ggplot2 R package (version 3.3.5) was used for data visualization. Venn diagrams were generated using either ssvFeatureEuler (seqsetvis R package, version 1.8.0) or ggVennDiagram (ggVennDiagram R package, version 1.1.4). Heatmaps were generated using pheatmap (pheatmap R package, version 1.0.12).

Results

CSF Proteomics Can Reliably Detect Brain Malignancy

We performed unbiased proteomic profiling of 73 CSF samples obtained from patients (mean age 63 years, 43.8% female) diagnosed with normal pressure hydrocephalus (NPH, n = 20), glioblastoma (GBM, n = 22), brain metastasis (BM, n = 17), or primary central nervous system lymphomas (CNSL, n = 14) (Figure 1, Table 1 and Supplementary Table S1). BMs were secondary to non-small cell lung cancer (NSCLC; lung) and invasive ductal carcinomas (breast) in 35.3% and 52.9% of cases, and all primary CNSL samples were diffuse large B-cell lymphomas.

For quality control (QC), all samples were spiked with S. cerevisiae invertase 2 (SUC2, internal control) and indexed retention time (iRT) peptides (chromatography control). We demonstrated limited variation in SUC2 intensities (Supplementary Figure S1A), and consistent iRT peptide elution profiles, thereby indicating reliable liquid chromatography (LC) performance (Supplementary Figure S1B). To ensure the reliability of our data, comprehensive QC samples were run after a fixed number of biological samples. The pooled QC samples were highly reproducible within processing ($R^2 = 0.92$) and technical ($R^2 = 0.93$) replicates (Supplementary Figure S1C). Correlation analysis within biological and random groups further confirmed the quality of our data (Supplementary Figure S1C). Using this proteomic workflow, we recovered 1333 unique proteins and showed that CNSL and BM (but not GBM) samples had a significantly more diverse proteome than NPH samples (Supplementary Figure S1D; $P = .0016$ and .039, respectively).

Next, we applied a 70% intra-group coverage threshold to obtain a dataset of 755 proteins for downstream analyses, representing an improvement in proteome coverage over the 506 proteins recovered by Schmid and colleagues in a comparable cohort using an identical coverage threshold (Supplementary Figure S2A–D). A difference between our study and Schmid et al. was the site of CSF samples; whereas we sampled CSF from intracranially-implanted CSF (Ommaya) reservoir, Schmid et al. obtained CSF using lumbar punctures. Tumor suppressor genes

---

**Figure 1.** Proteomic characterization of CSF from brain neoplasm patients. (A) Schematic representation of study workflow. (B) UMAP showing 2D representation of proteomic profiles, with each small node representing individual sample and large nodes representing median position of each patient cohort. BM, brain metastasis; CNSL, primary central nervous system lymphoma; FPR, false positive rate; GBM, glioblastoma; LC-MSMS, liquid chromatography with tandem mass spectrometry; NPH, normal pressure hydrocephalus; TPR, true positive rate; UMAP, uniform manifold approximation and projection.
eg, PTEN, NF1, TP53) and oncogenes (eg, EGFR, BRAF, PDGFRA) implicated in GBM, BM and CNSL were not reliably recovered by our protocol, which we attributed to our approach preferentially detecting secreted, rather than intracellular or membrane-bound proteins (Supplementary Table S1). Importantly, UMAP representation of CSF proteomes revealed coherent separation of each diagnostic group, suggesting that CSF proteomics are suitable for detecting malignancy, and discriminating between different brain neoplasms (Figure 1B).

To characterize malignancy-associated CSF biology, we performed differential protein analyses between NPH (non-malignant) and each brain neoplasm cohort (malignant) and identified 55 and 112 proteins that were consistently over- or under-represented in malignant CSF samples, respectively (Figure 2A, B; Supplementary Table S2). Functional annotation revealed that malignancy was associated with apoptotic signaling, glycolysis and heme metabolism, whereas non-malignancy was associated with elastic fibers- and extracellular matrix (ECM)-associated proteins, and neuronal and glial processes (Figure 2C, D; Supplementary Figure S3A).

Importantly, using published proteomic CSF data from GBM, BM, CNSL, Alzheimer’s disease (AD), Amyotrophic lateral sclerosis (ALS), and clinically-isolated syndrome of demyelination (CIS; first attack of multiple sclerosis (MS)),26 we demonstrated that our malignancy and non-malignancy signatures are significantly more deregulated in neoplastic disease (GBM, BM, CNSL) than non-neoplastic disease (AD, ALS, CIS/MS) (Figure 2E, F).

We utilized a logistic regression (LR) machine learning classifier framework to evaluate if our proteomic signatures can detect malignancy using CSF proteomics. We found that proteomic-based LR classifiers identified malignancy with a median AUROC of 0.94 (95% CI, 0.85–1.0), and this estimate ranged from 0.95 to 1.0 when evaluated for single neoplastic entities, demonstrating minimal neoplasm-specific bias (Figure 2G, Supplementary Table S3; Model 1). These performance metrics were reproduced using L1-regularized LR classifiers (Supplementary Table S3; Model 2), and feature selection using the Lasso procedure independently identified SERPINA3, HIST1H1E;HIST1H1D, IGFALS, HBA1, APOC2, FSTL3, SH3BGR1L3, FCGR3A, DPSS, HLA-C, IGHM, HYOU1, IGHV1-2 and GNPTG as malignancy-associated protein, and GALNT2, PI16, AGA, COCH, CCL14, PLXDC1, IGHV1-2 and GNPTG as non-malignancy-associated protein. Despite only 38% and 80% of proteins belonging to the malignant and non-malignant signatures being recovered in the Schmid cohort, respectively (Supplementary Figure S3B), malignancy in this

| Table 1. Patient and sample characteristics |
|--------------------------------------------|
| Description                               | Statistic | NPH | GBM | BM  | CNSL* | Total |
|--------------------------------------------|-----------|-----|-----|-----|-------|-------|
| Patient characteristics                     |           |     |     |     |       |       |
| Count                                      | n         | 20  | 22  | 17  | 14    | 73    |
| Sex %                                      |           |     |     |     |       |       |
| Age, years                                 | Mean (range) | 73 (48, 91) | 54 (24, 74) | 59 (42, 84) | 68 (57, 86) | 63 (25, 91) |
| Survival, days                             | Median |     |     |     |       |       |
| Treated %                                  |           | 95.5| 88.2| 78.6| 88.7**|       |
| Cell count, cells/µL                      | Mean (sd) | 0 (0) | 18.3 (30.6) | 33.4 (65.9) | 23.6 (57.9) | 25.0 (51.6) |
| Protein content, mg/dL                    | Mean (sd) | n.d. | 61.7 (56.7) | 70.0 (66.5) | 73.1 (86.5) | 67.5 (67.4)**|
| Malignant cytology %                      | Positive | n.a. | 19.0 | 70.6 | 25.0 | 38.0**|
| Tumor burden, cm³                         | Mean (sd) | n.a. | 14.8 (19.4) | 2.4 (4.6) | 19.2 (175) | 12.0 (16.8)**|
| Leptomeningeal %                          | Positive | n.a. | 63.6 | 64.7 | 42.8 | 58.5**|
| Touching ventricle %                      | Positive | n.a. | 63.6 | 23.5 | 85.7 | 56.6**|
| IDH status %                               | WT        | n.a. | 94.1*** | n.a. | n.a. | n.a. |
| MGMT status %                              | Methylated | n.a. | 33.3*** | n.a. | n.a. | n.a. |
| Lung % BMs                                 | n.a. |     |     | 35.3 | n.a. | n.a. |
| Breast % BMs                               | n.a. |     |     | 52.9 | n.a. | n.a. |
| Other % BMs                                | n.a. |     |     | 11.8 | n.a. | n.a. |

n.a., not applicable; n.d., not determined; sd, standard deviation.

*All CNSL cases are primary CNSL (PCNSL) neoplasms (diffuse large B-cell lymphoma).

**Only neoplastic samples (GBM, BM and CNSL) included in calculation.

***IDH and MGMT status determined for 17/22 and 18/22 GBM samples, respectively.
Elastic fibres/ECM (Figure S3C–H and Table S3).

validity of our signatures (Figure 2H, Supplementary Table S3). Demonstrating the external
independent cohort was classified with a median AUROC of 0.82 (95% CI, 0.72–0.91), demonstrating the external

validity of our signatures (Figure 2H, Supplementary Figure S3C–H and Table S3).

Upon consideration of clinical covariates, including age, sex, presence of malignant cytology in CSF, and leptomeningeal status, we observed a negligible 0–0.02 improvement in AUROC performance (Supplementary Figure S3D,
Having established that the upper-bound AUROC performance of CSF proteomics ranges from 0.95 to 1.0, we next sought to identify focused biomarker panels. We applied our machine learning-based framework to nominate individual proteins for each type of brain neoplasm (Figure 3A; Supplementary Figure S6A and Tables S4–5). Using an orthogonal approach, we also identified candidate biomarkers using Lasso-based feature selection but found this list of proteins to be redundant with the approach described above. Next, we evaluated combinatorial classifiers, termed “triplex classifiers,” that comprised of three-way combinations of the top proteins identified by our uniplex classifiers, and these classifiers were AUROC-ranked to identify the top combinations of biomarkers (Figure 3A; Supplementary Figure S6B; Table S5). Among the top 25 triplex classifiers, we focused on further characterizing GAP43 (GBM-specific), TFF3 (BM-specific), and CACNA2D2 (CNSL-specific) (Supplementary Figure 3B). Whereas GAP43, TFF3 and CACNA2D2 uniplex classifiers had a median AUROC of 0.69, 0.75 and 0.84, in classifying GBM, BM and CNSL, respectively, the combined triplex classifier performed better with a median AUROC of 0.75, 0.88 and 0.89, respectively (Figure 3C, Supplementary Table S4). We also evaluated these uniplex and triplex classifiers on the subset of IDH-wt GBM (16/22 glioma samples) and demonstrated AUROCs of 0.68 (95% CI, 0.465–0.87, \( P = .04 \)) and 0.75 (95% CI, 0.58–0.95, \( P = .01 \)), respectively.

To explore the biology of these biomarkers, we assessed their transcriptional profiles using publicly-available single cell transcriptomic (scRNAseq) profiles of GBM,\(^{29}\) lung adenocarcinoma BM,\(^{28}\) and CNSL (Supplementary Figure S7).\(^{31}\) In GBM scRNAseq data, GAP43 expression was largely restricted to GBM cells, limited in lymphoid and myeloid cells, and entirely absent in BM and CNSL cells. TFF3 was exclusively expressed in BM. CACNA2D2 expression was absent in GBM and CNSL and detected in a subpopulation of BM tumor cells. Collectively these data suggest that GAP43 and TFF3 observed in CSF are derived from tumor populations, unlike other candidate biomarkers (eg, SOD2, PZP, and CTSZ) that exhibit non-specific patterns of expression. The lack of transcriptomic CACNA2D2 expression in CNSL samples suggests a non-neoplastic source, or lack of correlation between transcriptomic and proteomic expression.

### Pathway-Level Analyses Identify Survival Associations

In our cohort of brain tumor patients, BM and CNSL patients had similar survival rates (564 and 649 median days, respectively), whereas GBM patients had a median survival of 2246 days (Figure 4A). To explore the biology associated with survival, we evaluated seven diverse cancer-associated signature sets in our cohort, including pan-cancer and GBM-specific signatures from the Clinical Proteomic Tumor Analysis Consortium (CTPAC).\(^{20,28,47-51}\) In addition to these signatures discriminating malignant from non-malignant samples (Supplementary Figure S8A), survival analyses revealed that mesenchymal and invasion-associated signatures trended towards unfavorable...
survival outcomes (Supplementary Figure S8B). Our small sample size limited us from identifying novel prognostic biomarkers (Supplementary Figures S9A–D and S10A). To overcome this limitation, we undertook a pathway-level analysis which validated previous findings reported by Schmid et al.16 (Figure 4B–D, Supplementary Figures S9E and S10B–E). Specifically, we demonstrated that markers associated with blood–brain barrier disruption [eg, blood coagulation (C4BPB, COL1A1, CPB2, F10, F13B, F2, F9, FGA, FGG, PLG, PROS1, PROZ, SERPINA1, SERPINA5, SERPINC1, SERPIND1, VWF)], and complement pathway activation (C2, C3, C4B, C4BPA, C4BPB, C5, C6A, C6B, C9, CFB, CFH, CFHR1, CFI, CPB2, CPN1, CPN2, F2, SERPING1, VTN)] (Figure 4B), angiogenesis (APOE, APOH, CDH5, HRG, PROC, SEMA6A, THBS1) (Figure 4C) and stemness (KIT, NOTCH2) (Figure 4D) were associated with unfavorable survival outcomes.

Discussion

In this proof-of-concept study, we show the feasibility of high protein recovery and comprehensive proteomic analyses on low sample volumes, using the MStern sample processing approach.52 We recovered 755 proteins with as little as 30 µL of CSF per sample. Along with clearly distinguishing malignant from non-malignant samples, we were able to build and externally validate classifiers for distinguishing GBM, BM, and CNSL. With an eye toward relevance to clinical practice, we further developed a tripleplex classifier comprised of GAP43, TFF3, and CACNA2D2 that can together distinguish between the three disease entities. Prospective validation of these findings can have a profound impact on longitudinal surveillance of these malignant entities.
Using CSF based proteomics, we were able to distinguish malignant from non-malignant samples and identified pathway-level alterations of prognostic significance. These were both externally validated using data from the Schmid et al. cohort. While CSF proteomics offered similar classification specificity to CSF cytology alone, we demonstrated that CSF proteomics outperformed CSF cytology with respect to sensitivity, thereby representing a significant value-add for distinguishing malignant from non-malignant samples. We were also able to develop machine learning-based classifiers to reliably distinguish between the three brain tumor entities; this was again externally validated using the Schmid cohort. Since these classifiers depended on larger protein panels, traditional antibody-driven methods, such as Western Blotting and ELISA, are inefficient and represent a potential bottleneck. Selective reaction monitoring (SRM) and, more recent, parallel reaction monitoring (PRM) utilize stable isotope-labeled peptides as internal standards of previously detected candidates. Targeted proteomics assays are quantitative tools that enable the robust, sensitive and multiplexable quantitation of proteins without the need of antibodies, but require significant time for initial development, validation and implementation of these assays. Thus, recognizing that large-scale mass spectrometric analyses can be challenging in the routine clinical setting, we further built on our approach by identifying the top 3 biomarker proteins that can together distinguish between GBM, BM, and CNSL with high precision (AUROC 0.75–0.89). Further prospective validation of these biomarkers will inevitably warrant consideration of upfront implantation of CSF reservoirs for routine monitoring of the tumor microenvironment during therapy and surveillance, through scheduled sampling of the CSF to assess for relevant biomarkers. Currently, such strategies have rarely been implemented. In 2016, Brown et al. used CSF reservoirs for the delivery of CART cell therapy, with regular monitoring of cytokines and immune cells within the CSF, in a patient with multifocal recurrence of GBM with remarkable success. This was part of a Phase I trial that is currently underway (NCT02208362).

The Trefoil Factor Family (TFF1, TFF2, and TFF3) are proteins secreted by normal mucous secretory epithelia. TFF3 transcription and translation has been reported in various cancers. The oncogenic behavior of TFF3 is mediated through promotion of cancer cell migration and invasion, along with down regulation of apoptotic signaling. In breast cancer, its expression has been associated with lymph node involvement and increased propensity for local metastasis. Of note, its mRNA expression has been linked to higher likelihood of breast cancer metastases to the CSF and bones. Other cancers in which TFF3 has been associated with include gastric, pancreatic, colorectal, cervical, and prostate. As such, our work here contributes to a growing body of evidence that TFF3 is a pan-cancer biomarker of metastasis, and future studies validating TFF as a CSF biomarker of brain metastases are warranted.

We found a significant association between CNSL and CACNA2D2 expression, a voltage-gated calcium channel receptor found in numerous tissues including the brain. The role of this protein in cancer is unclear, with conflicting data on whether it is a tumor suppressor or oncogenic protein. While CACNA2D2 could have simply been shed from normal brain tissue, its specific association with CNSL warrants exploration. This is particularly relevant given the emerging role of extracellular vesicles (EVs), including exosomes, in cancer biology. EVs are cell-derived proteins secreted by normal mucous secretory epithelia. TFF3 transcription and translation has been reported in various cancers. The oncogenic behavior of TFF3 is mediated through promotion of cancer cell migration and invasion, along with down regulation of apoptotic signaling. In breast cancer, its expression has been associated with lymph node involvement and increased propensity for local metastasis. Of note, its mRNA expression has been linked to higher likelihood of breast cancer metastases to the CSF and bones. Other cancers in which TFF3 has been associated with include gastric, pancreatic, colorectal, cervical, and prostate. As such, our work here contributes to a growing body of evidence that TFF3 is a pan-cancer biomarker of metastasis, and future studies validating TFF as a CSF biomarker of brain metastases are warranted.
vesicles released by all cells, and their protective lipid membranes permits protected travel throughout the CSF and blood stream in the body. Complex interactions by cancer cells with other cells in the tumor microenvironment can be mediated through EVs, via exchange of biologic factors including DNA, RNA, and proteins.

Limitations

Our findings should be interpreted with several caveats, including the retrospective nature of the study precluding comprehensive correlative analyses with clinical data. Small sample size and overfitting were also a concern; however, we used several strategies to minimize the risk of spurious findings, including L1-normalization, leave-one-out cross-validation and evaluation of models using data from the Schmid cohort. Furthermore, given the retrospective acquisition of samples, certain sample characteristics were unavailable, including IDH status in 5/22 glioma samples. Similarly, most gliomas included in this study were classified prior to the updated 2021 WHO guidelines, thus it is possible that some glioma samples in which IDH status was not determined are not grade IV GBM. Within the BM cohort, it was not documented whether CSF was acquired while the primary tumor still existed or not, and this could represent a potential confounding variable. The exquisitely long survival of our GBM cohort (median of 2246 days), a selection of ultra-long survivors by virtue of our clinical protocol for implantation of CSF reservoirs in GBM patients, also necessitates the cautious interpretation of our data pertaining to this tumor entity. However, despite the longer survival times in our GBM cohort, the GBM classifier validated similarly to the BM- and CNSL-classifiers when applied to the independent Schmid cohort dataset. Although our triplex classifier presents the opportunity for a clinically facile approach to reliably distinguishing between the three cancer entities and potentially surveillance for tumor recurrence, neither of GAP43, TFF3 or CACNA2D2 were recovered in the Schmid cohort, thereby precluding evaluation of their external validity in an independent cohort. The limited biomarker recovery in the Schmid cohort should not serve as an indication of protein instability of insufficiency, but rather indicate the need for future validation using targeted assays (eg, SRM, ELISA), which are known to be more sensitive. Although corroboration with tumor-derived scRNAseq data offered exciting insight into the potential source of proteins detected in our CSF samples, it must be emphasized that scRNAseq data were derived from tumor tissue and not CSF samples. Similarly, signatures and associations identified in this study are restricted to free proteins shed into the CSF and do not reflect a comprehensive snapshot of the composition of the tumor microenvironment or intracellular protein contents that may offer more tumor-specific biology. One strategy to address this is to profile CSF-derived extracellular vesicles (EVs), reasoning that EV contents are more reflective of intracellular proteins; however, this comes with the challenge of requiring larger CSF volumes and extensive sample preparation that precludes scalability. Our current analysis does not account for alternate splice site variations and post-translational modifications. In addition to correlative analyses with matched tumor and plasma, analyses of exosome cargo and incorporation of other -omic analyses, including phosphoproteomics, glycoproteomics, metabolomics, genomics, and lipidomics will be critical. Despite these limitations, our study is the first of its kind to demonstrate the feasibility of high protein recovery from very limited CSF sample, yielding promising biomarkers for disease surveillance and prognosis.

Future Directions

Given the retrospective nature of the current study, prospective validation of the proteomic signatures identified will be required. In parallel, development and optimization ELISA and PRM-based assays for small- (eg, triplex classifiers) and large-panel protein signatures, respectively, will facilitate the translation of the classifiers into a clinical setting. In particular, PRM enables targeted quantification of tens to hundreds of proteins, and while is primarily used as a research tool, its scalability, cost-effectiveness, and information-rich readouts make it a promising tool for future clinical applications. Finally, while our glioma cohort was predominantly comprised of IDH-wt GBM patients, we recognize the value of using a CSF proteomic approach to discriminate between different glioma entities, including IDH-mutant gliomas, and future studies addressing this will be required.

Supplementary material

Supplementary material is available online at Neuro-Oncology Advances online.

Keywords

brain metastasis | biomarkers | cerebrospinal fluid | disease surveillance | glioblastoma

Funding

This research work was supported by the 2020 William Donald Nash Brain Tumor Research Fellowship awarded to NM. This work was partially supported by a Canadian Institute of Health Research Project Grant (PJT154357) to TK. We thank Andrew Macklin for technical assistance. This research was funded in part by the Ontario Ministry of Health and Long-Term Care. TK was supported through the Canadian Research Chair program.

Competing Interests

The research was conducted in the absence of any commercial/financial relationships that could be construed as a conflict of interest.
Declarations

**Ethics Approvals and Consent to Participate:** All participants provided informed consent for this Biorepository study. Approval for this study was granted by the Penn State IRB (#2914). Specimens and associated data were released to the protocol “Molecular testing of nervous system cancers as a classification tool” following approval of the Biorepository Data Oversight Committee (#20-0002). Data was released in a de-identified manner via an honest broker system. The patient’s privacy was protected in accordance with both Penn State’s IRB and HIPAA guidelines.

**Consent for Publication:** All participant information included in the current study is unidentifiable.

Author Contributions

Study conception and design: TK, AM; Acquisition of experimental data: MT, AS, SK, MG, TK, AM; Analysis and interpretation of data: NM, SK, TK, AM; Drafting of Manuscript: NM, SK, AM; Supervision: TK, AM.

Data Availability

Proteomic data have been deposited to MassIVE (identifier: MSV000089062). The corresponding FTP link is ftp://massive.ucsd.edu/MSV000089062/.

References

1. Ostrom QT, Cioffi G, Gittleman H, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2012–2016. *Neuro-oncology*. 2019;21(Supplement_5):v1–v100. doi:10.1093/neuonc/noz150. PMID: 31675094; PMCID: PMC6823730.
2. Stupp R, Hegi ME, Mason WP, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol*. 2009;10(5):459–466.
3. Stupp R, Taillibert S, Kanner A, et al. Effect of tumor-treating fields plus maintenance Temozolomide vs Maintenance Temozolomide alone on survival in patients with Glioblastoma: A Randomized Clinical Trial. *JAMA*. 2017;318(23):2306–2316.
4. O’Connell K, Romo CG, Grossman SA. Brain metastases as a first site of recurrence in patients receiving chemotherapy with controlled systemic cancer: a critical but under-recognized clinical scenario. *Curr Treat Options Neurol*. 2019;21(11):1–7.
5. Nieder C, Spanne O, Mehta MP, Grosu AL, Geinitz H. Presentation, patterns of care, and survival in patients with brain metastases: what has changed in the last 20 years? *Cancer*. 2011;117(11):2505–2512.
6. Andre F, Grunenwald D, Pujol J, et al. Patterns of relapse of NSCLC nonsmall-cell lung carcinoma patients treated with preoperative chemotherapy: Should prophylactic cranial irradiation be reconsidered? *Cancer*. 2001;91(12):2394–2400.
7. Freilich RJ, Seidman AD, Deangelis LM. Central nervous system progression of metastatic breast cancer in patients treated with paclitaxel. *Cancer*. 1995;76(2):232–236.
8. Malone H, Yang J, Hershman DL, et al. Complications following stereotactic needle biopsy of intracranial tumors. *World Neurosurg*. 2015;84(4):1084–1089.
9. Kucharczyk MJ, Parpia S, Whitton A, Greenspoon JN. Evaluation of pseudoprogression in patients with glioblastoma. *Neuro-oncol Pract*. 2017;4(2):120–134.
10. Zikou A, Sioka C, Alexiou GA, et al. Radiation necrosis, pseudoprogression, pseudoresponse, and tumor recurrence: imaging challenges for the evaluation of treated gliomas. *Contrast Media Mol Imaging*. 2018;2018:6828396. doi:10.1155/2018/6828396. PMID: 30627060; PMCID: PMC6305027.
11. Miller AM, Shah RH, Pentsova EI, et al. Tracking tumour evolution in glioma through liquid biopsies of cerebrospinal fluid. *Nature*. 2019;565(7741):654–658.
12. De Mattos-Arruda L, Mayor R, Ng CK, et al. Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nat Commun*. 2015;6:8839. doi:10.1038/ncomms8839. PMID: 26554728; PMCID: PMC5426516.
13. Newman AM, Bratman SV, To J, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med*. 2014;20(5):548–554.
14. Nassiri F, Chakravarty A, Feng S, et al. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes. *Nat Med*. 2020;26(7):1044–1047.
15. Shen SY, Singhania R, Fehringer G, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*. 2018;563(7732):579–583.
16. Schmid D, Warnken U, Latzer P, et al. Diagnostic biomarkers from proteomic characterization of cerebrospinal fluid in patients with brain malignancies. *J Neurochem*. 2021;158(2):522–538. doi:10.1111/jnc.15350. PMID: 33735443.
17. Zhang H, Liu T, Zhang Z, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*. 2016;166(3):755–765.
18. Mertins P, Mani D, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*. 2016;534(7605):55–62.
19. Oh S, Yeom J, Cho HJ, et al. Integrated pharmaco-proteogenomics defines two subgroups in isocitrate dehydrogenase wild-type glioblastoma with prognostic and therapeutic opportunities. *Nat Commun*. 2021;12(11):16.
20. Wang LB, Karpova A, Gritsenko MA, et al. Proteogenomic and metabolic characterization of human glioblastoma. *Cancer Cell*. 2021;39(4):509–528.e20.
21. Berger ST, Ahmed S, Muntel J, et al. MStern blotting-high throughput polyvinylidene fluoride (PVDF) membrane-based proteomic sample preparation for 96-well plates. *Mol Cell Proteom*. 2015;14(10):2814–2823.
22. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*. 2008;26(12):1367–1372.
23. Wojtowicz EE, Leechman ER, Hermans KG, et al. Ectopic miR-125a expression induces long-term repopulating stem cell capacity in mouse and human hematopoietic progenitors. *Cell Stem Cell*. 2016;19(3):383–396.
24. Bader JM, Geyer PE, Müller JB, et al. Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer’s disease. *Mol Syst Biol*. 2020;16(6):e9356.
Bereman MS, Beri J, Enders JR, Nash T. Machine learning reveals protein signatures in CSF and plasma fluids of clinical value for ALS. *Sci Rep.* 2018;8(1):1–14.

Stoop MF, Runia TF, Stigl C, et al. Decreased neuro-axonal proteins in CSF at first attack of suspected multiple sclerosis. *Proteom Clin Appl.* 2017;11(11–12):1700005.

Wei B, Liu Z, Fan Y, et al. Analysis of cellular heterogeneity in immune microenvironment of primary central nervous system lymphoma by single-cell sequencing. *Front Oncol.* 2021;11:683007. doi:10.3389/fonc.2021.683007. PMID: 34671548; PMCID: PMC8233033.

Nettel C, Laffly J, Filbin MG, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell.* 2019;178(4):835–849.e21.

Kim N, Kim HK, Lee K, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun.* 2020;11(1):1–15.

Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell. 2018;174(3):3573–3587.e29. doi:10.1016/j.cell.2018.04.048. PMID: 34062119; PMCID: PMC6283499.

Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell.* 2019;179:1–188–1902.e21.

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptionomic data across different conditions, technologies, and species. *Nat Biotechnol.* 2018;36(5):411–420.

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol.* 2015;33(5):495–502.

Todorov V, Todorov MV. Package “rroc”; 2016.

Koruszynski N, Nathan A, Millard N, Raychaudhuri S (2019) Presto scales Wilcoxon and auROC analyses to millions of observations. *BioRxiv.* 653253. preprint: not peer reviewed. doi:10.1101/653253.

Korotkevich G, Sukhov V, Budin N, et al. (2021) Fast gene set enrichment analysis. *BioRxiv.* 606012. preprint: not peer reviewed. doi:10.1101/606012.

Liberson A, Birger C, Thorvaldsdottir H, et al. The molecular signatures database hallmark gene set collection. *Cell Syst.* 2015;16:417–425.

Fabregat A, Jupe S, Matthews L, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 2018;46(D1):D649–D655.

Gelman A, Su Y-S, Yajima M, et al. Package “arm”: Data Analysis Using Regression and Multilevel/Hierarchical Models. 2015.

Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in *R*. *Bioinformatics.* 2005;21(20):3940–3941.

Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis of microarray and RNA-seq data. *BMC Bioinformatics.* 2013;14(1):1–15.

Kassambara A, Kosinski M, Biecek P, Fabian S. Package “survminer.” *Drawing Survival Curves using “ggplot2” (R package version 03 1).* 2017.

Bailey MH, Tokheim C, Porta-Pardo E, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell.* 2018;173(2):371–385. e18.

Sinha A, Hussain A, Ignatchenko V, et al. *N*-Glycoproteomics of patient-derived xenografts: a strategy to discover tumor-associated proteins in high-grade serous ovarian cancer. *Cell Syst.* 2019;8(4):345–351.e4.

Kim Y, Jeon J, Mejia S, et al. Targeted proteomics identifies liquid biopsy signatures for extracapsular prostate cancer. *Nat Commun.* 2016;7(1):11906.

Guldbrandsen A, Lereim RR, Jacobsen M, et al. Development of robust targeted proteomics assays for cerebrospinal fluid biomarkers in multiple sclerosis. *Clin Proteom.* 2020;17(1):1–21.

Chen F, Chandrashekar DS, Varambally S, Creighton CJ. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. *Nat Commun.* 2019;10(1):5679.

Richards LM, Whiteley OKN, MacLeod G, et al. Gradient of developmental and injury response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity. *Nat Cancer.* 2021;2(2):157–173.

Verhaak RGB, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in IDH1, IDH2, EGFR, and NF1. *Cancer Cell.* 2010;17(1):98–110.

Nimel AJ, Ragan T, Shih BB, et al. Immune cell gene signatures for profiling the microenvironment of solid tumors. *Cancer Immunol Res.* 2018;6(11):1386–1400.

Yuan H, Yan M, Zhang G, et al. CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.* 2019;47(D1):D900–D908.

Potriquet J, Lahováriová M, Bethony JM, Mulvenna J. A modified FASP protocol for high-throughput preparation of protein samples for mass spectrometry. *PloS One.* 2017;12(7):e0175867.

Ronsein GE, Pamir N, von Haller PD, et al. Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics. *J Proteome.* 2015;113:388–399. doi:10.1016/j.jprot.2014.01.017. PMID: 25449833; PMCID: PMC4259393.

Mao J, Wang J, Zhang W, et al. Convenient multicomponent reaction synthesis of novel pyrano [4, 3-b] pyran derivatives via a domino reaction under microwave irradiation. *Arkivoc.* 2016;3:171–186. doi:10.3998/ark.5550190.p039.398.

Brown CE, Alizadeh D, Starr R, et al. Regression of glioblastoma after chimeric antigen receptor T-cell therapy. *N Engl J Med.* 2016;375(26):2561–2569. doi:10.1056/NEJMoa1610497. PMID: 28029927; PMCID: PMC5330684.

Samson MH, Chajariarit P, Nortvig H, et al. Trefoil factor family peptides in human saliva and cyclical cervical mucus. Method evaluation and results on healthy individuals. *Clin Chem Lab Med.* 2011;49(5):861–868.

Hoffmann W. Trefoil factors. *Cell Mol Life Sci CMLS.* 2005;62(24):2923–2938.

Chen Y-H, Lu Y, De Plaen IG, Wang L-Y, Tan X-D. Transcription factor NF-kB signals anatoikic function of trefoil factor 3 on intestinal epithelial cells. *Biochem Biophys Res Commun.* 2000;274(3):576–582.

Kinoshita K, Taupin DR, Itoh H, Podolsky DK. Distinct pathways of cell migration and antiapoptotic response to epithelial injury: structure-function analysis of intestinal human trefoil factor. *Mot Cell Biol.* 2000;13:4680–4690.

Smid M, Wang Y, Klijn JG, et al. Genes associated with breast cancer metastatic to bone. *J Clin Oncol.* 2006;24(15):2261–2267.

Sharif AH, Griffiths AB, Tilby MT, Westley BR, May FE. TF3F is a normal breast epithelial protein and is associated with differentiated phenotype in early breast cancer but predisposes to invasion and metastasis in advanced disease. *Am J Pathol.* 2012;180(3):904–916.

Magbanua MJM, Melisko M, Roy R, et al. Molecular profiling of tumor cells in cerebrospinal fluid and matched primary tumors from metastatic breast cancer patients with leptomeningeal carcinomatosis. *Cancer Res.* 2013;73(23):7134–7143.

Taupin D, Pedersen J, Famiari M, et al. Augmented intestinal trefoil factor (TF3F) and loss of pS2 (TF1F) expression precedes metastatic differentiation of gastric epithelium. *Lab Invest.* 2001;81(3):397–408.

Terris B, Blavari E, Cnogorac-Jurcovic T, et al. Characterization of gene expression profiles in intraductal papillary-mucinous tumors of the pancreas. *Am J Pathol.* 2002;160(5):1745–1754.

Taupin D, Döi K, Yeomans N, Giraud A. Conserved expression of intestinal trefoil factor in the human colonic adenoma-carcinoma sequence. *Lab Invest.* 1996;75(1):25–32.

Yuan Z, Chen D, Chen X, Yang H, Wei Y. Overexpression of trefoil factor 3 (TF3F) contributes to the malignant progression in cervical cancer cells. *Cancer Cell Int.* 2017;17(1):1–13.

Faith DA, Isaacs WB, Morgan JD, et al. Trefoil factor 3 overexpression in postactic carcinoma: prognostic importance using tissue microarrays. *Prostate.* 2004;61(3):215–227.
68. Fiske JL, Fomin VP, Brown ML, Duncan RL, Sikes RA. Voltage-sensitive ion channels and cancer. Cancer Metastasis Rev. 2006;25(3):493–500.
69. Kunzelmann K. Ion channels and cancer. J Membr Biol. 2005;205(3):159–173.
70. Lerman MI, Minna JD. The 630-kb lung cancer homozygous deletion region on human chromosome 3p21. 3. Identification and evaluation of the resident candidate tumor suppressor genes. Cancer Res. 2000;60(21):6116–6133.
71. Gao B, Sekido Y, Maximov A, et al. Functional properties of a new voltage-dependent calcium channel α2δ auxiliary subunit gene (CACNA2D2). J Biol Chem. 2000;275(16):12237–12242.
72. Zhang Y, Liu Y, Liu H, Tang WH. Exosomes: biogenesis, biologic function and clinical potential. Cell Biosci. 2019;9(1):19. doi:10.1186/s13578-019-0282-2. PMID: 30815248; PMCID: PMC6377728.
73. Hornung S, Dutta S, Bitan G. CNS-derived blood exosomes as a promising source of biomarkers: opportunities and challenges. Front Mol Neurosci. 2020;13.
74. Picotti P, Aebersold R. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. Nat Methods. 2012;9(6):555–566. doi:10.1038/nmeth.2015. PMID: 22689653.
75. Gramolini A, Peterman S, Kislinger T. Mass spectrometry–based proteomics: a useful tool for biomarker discovery? Clin Pharmacol Ther. 2008;83(5):758–760.
76. Shi T, Song E, Nie S, et al. Advances in targeted proteomics and applications to biomedical research. Proteomics. 2016;16(15–16):2160–2182.