PLOS ONE

RESEARCH ARTICLE

Investigating centrifugal filtration of serum-based FTIR spectroscopy for the stratification of brain tumours

Ashton G. Theakstone[1], Paul M. Brennan[2], Michael D. Jenkinson[3,4], Royston Goodacre[5], Matthew J. Baker[6,7]*

1 Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, United Kingdom, **2** Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom, **3** The Walton Centre NHS Foundation Trust, Liverpool, United Kingdom, **4** Department of Pharmacology & Therapeutics, University of Liverpool, Liverpool, United Kingdom, **5** Department of Biochemistry and Systems Biology, University of Liverpool, Liverpool, United Kingdom, **6** Dxcover Limited, Glasgow, United Kingdom, **7** Faculty of Clinical and Biomedical Sciences, University of Central Lancashire, Preston, United Kingdom

* matthew.baker@dxcover.com, MBaker10@uclan.ac.uk

Check for updates

OPEN ACCESS

Citation: Theakstone AG, Brennan PM, Jenkinson MD, Goodacre R, Baker MJ (2023) Investigating centrifugal filtration of serum-based FTIR spectroscopy for the stratification of brain tumours. PLoS ONE 18(2): e0279669. https://doi.org/10.1371/journal.pone.0279669

Editor: Tommaso Lomonaco, University of Pisa, ITALY

Received: December 12, 2022

Accepted: January 31, 2023

Published: February 17, 2023

Copyright: © 2023 Theakstone et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This research was funded by Cancer Research UK (Grant number A28345).

Competing interests: Author MJB is a director of Dxcover Limited. All other authors declare no conflict of interest. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Abstract

Discrimination of brain cancer *versus* non-cancer patients using serum-based attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy diagnostics was first developed by Hands *et al* with a reported sensitivity of 92.8% and specificity of 91.5%. Cameron *et al*. then went on to stratifying between specific brain tumour types: glioblastoma multiforme (GBM) vs. primary cerebral lymphoma with a sensitivity of 90.1% and specificity of 86.3%. Expanding on these studies, 30 GBM, 30 lymphoma and 30 non-cancer patients were selected to investigate the influence on test performance by focusing on specific molecular weight regions of the patient serum. Membrane filters with molecular weight cut offs of 100 kDa, 50 kDa, 30 kDa, 10 kDa and 3 kDa were purchased in order to remove the most abundant high molecular weight components. Three groups were classified using both partial least squares-discriminate analysis (PLS-DA) and random forest (RF) machine learning algorithms; GBM *versus* non-cancer, lymphoma *versus* non-cancer and GBM *versus* lymphoma. For all groups, once the serum was filtered the sensitivity, specificity and overall balanced accuracies decreased. This illustrates that the high molecular weight components are required for discrimination between cancer and non-cancer as well as between tumour types. From a clinical application point of view, this is preferable as less sample preparation is required.

Introduction

Brain cancer diagnosis is challenging. The most common symptoms are non-specific (such as headaches) and are more likely to be associated with a non-tumour diagnosis [1–3]. As many as two thirds of patients are diagnosed in the Emergency Department when their symptoms have deteriorated, with the majority of these patients having previously visited their primary care doctor multiple times [4]. There is a need for a rapid, cost-effective and non-invasive tool for earlier diagnosis.

A vibrational spectroscopic technique, attenuated total reflection (ATR) Fourier transform infrared (FTIR) spectroscopy, has been applied to earlier detection and diagnosis of brain tumours [5, 6]. FTIR spectroscopy involves irradiating samples with infrared light where the absorbance of light results in an IR spectrum that is representative of specific components within the sample. Indication of disease states is possible through imbalances of biomolecular components and diagnostic outputs are achievable with machine learning algorithms [7]. ATR-FTIR mode uses an internal reflection element (IRE) where an evanescent wave extends beyond the IRE and penetrates the sample that is in direct contact [8].

Serum-based ATR-FTIR combined with machine learning algorithms can reliably predict which patients with symptoms of a possible brain tumour actually have a tumour on brain imaging. Hands *et al*. were the first to investigate the use of serum for ATR-FTIR spectroscopic analysis for brain tumour diagnosis, comparing brain tumour and asymptomatic non-tumour patients. Subsequent studies have included symptomatic non-tumour patients as well as investigating predictions of tumour grade and subtype [9–11]. The earlier work used a traditional, time-consuming ATR-FTIR set-up with a fixed-point diamond IRE. A newer, high-throughput approach uses silicon-based IRE (SIRE) sample slides. These SIREs are disposable and have multiple sampling points, which allows for high-throughput and batch processing [12, 13].

With this technique, brain tumours can be detected with a sensitivity of 88.7%, specificity of 94.7% and overall balanced accuracy of 91.7% [14]. To further improve test performance, we investigated whether specific molecular weight regions of patient serum improved detection and stratification. Blood serum contains over 20,000 different proteins with a wide range of molecular weights, dominated by human serum albumin (HSA); 30–50 g/L is considered normal [15]. Imbalances within protein concentrations in serum may relate to specific disease states and the low molecular weight fraction of serum may contain cancer-specific diagnostic information [16, 17].

Commercially available centrifugal filters can fractionate serum according to a molecular threshold, and so aid investigation of specific molecular weight fractions of serum. Traditionally, these filters are used to separate and remove rapidly the most abundant high molecular weight proteins from the less abundant low molecular weight molecules (*viz*. metabolites). One concern with these filters is the extra sample preparation required, the binding of small molecules to proteins which are then removed by filtrations. The reported potential for contamination from the filter membrane, has been resolved by Bonnier *et al*. who developed a centrifugal washing technique to remove any trace glycerine from the filter membranes [18, 19].

Here, we use the serum-based ATR-FTIR technique to investigate six (five fractions plus unfiltered whole serum) different molecular weight regions of serum for the stratification of brain cancer patients against non-cancer controls (Hands *et.al* previously reported sensitivity of 92.8% and specificity of 91.5% [9]). We also explore the stratification between tumour types; GBM and primary cerebral lymphoma (Cameron *et al*. previously reported sensitivity of 90.1% and specificity of 86.3% [12]).

Analysis of the patient serum corresponded with previous published work and involved an unsupervised exploratory principal component analysis (PCA) followed by supervised machine learning methods including random forest (RF) and partial least squares-discriminant analysis (PLS-DA). PCA involves an orthogonal linear transformation of the data to determine any separation between the classes. Any variance can be displayed within a scores plot as principal components (PC) with the first PC responsible for the greatest variance [20]. RF and PLS-DA are supervised classification algorithms where RF uses a Classification and Regression Trees (CART) technique to build an ensemble of decision trees as independent models and predictions are based on a majority vote within the forest [21, 22]. PLS-DA combines PLS regression and linear discriminant analysis to reduce the dimensionality of complex

data to revel hidden patterns. In binary classifications the technique separates classes by dividing the data space into two distinct regions and new variables are formed called PLS components, with the first PLS component accounting for the greatest variance (PLS1). The corresponding loadings plots can further explain the variance by highlighting the regions where highest disparity between the classes is observed [23, 24].

Materials and methods

Patient serum samples (n = 90) were obtained from the Walton Centre NHS Trust (Liverpool, UK) and the Royal Preston Hospital (Preston, UK) with informed written consent, under Ethics approval code (Walton Research Bank BTNW/WRTB 13_01/BTNW Application #1108). Included within the study were 30 glioblastoma (GBM) patients, 30 primary cerebral lymphoma patients and 30 asymptomatic control patients.

The patient serum was fractionated sequentially through five different size molecular weight filters (100 kDa, 50 kDa, 30 kDa, 10 kDa and 3 kDa) (Amicon Ultra-0.5 mL, Merck, Germany). The samples were centrifuged at 14,000 x*g* for 30 min to collect molecular weight fractions. The filtrate was collected and analysed so each portion represented the molecular weights less than the cut-off point (E.g., <100 kDa). This resulted in 6 serum samples per patient (including unfiltered whole serum), with a total of 540 samples.

Before the filters were used, they were centrifugally washed with 0.1M NaOH and MilliQ water through the following steps; 30 min with 0.1M NaOH at 14,000 x*g*, followed by 2 times 30 min with MilliQ water at 14,000 x*g*, and finally 2 min upside down at 1,000 x*g* to remove any remaining liquid. The washing was necessary to remove any residual glycerine coating on the ultrafiltration membranes as indicated by the manufacturer, to ensure no interferences within the sample spectra. Within the S1 File there is example serum spectra illustrating both washed and unwashed filters, highlighting the need for the pre-analytical washing steps.

Patient serum, either whole or molecular weight fraction (3 μL), was deposited onto a SIRE optical sample slide (Dxcover Ltd, Glasgow, UK) and air dried before spectroscopic data collection. All serum spectra were collected on a Perkin Elmer Spectrum 2 FTIR spectrometer (Perkin Elmer, London, UK), utilising a Specac Quest ATR accessory unit with a specular reflectance puck (Specac Ltd., London, UK), allowing a Dxcover optical sample SIRE (Dxcover Ltd., Glasgow, UK) to be placed directly on top of the aperture. Each sample SIRE contains four wells where one remains blank as the background and the other three were used as sample repeats, with each three wells analysed three times. Nine spectra per patient were collected within the range of 4000–450 cm^{-1}, at a resolution of 4 cm^{-1}, with 1 cm^{-1} data spacing and 16 co-added scans; resulting in a total of 4,860 spectra acquired. The typical time for spectral collection was 15 min per patient sample slide (9 repeats and background).

The spectroscopic data analysis was completed using the R Statistical Computing Environment, MATLAB R2020a software with the PRFFECT toolbox [25] or a PCA code written in house. Data pre-processing was applied to reduce computational burden and improve classification algorithms. The techniques used match previous published work including a min-max normalisation, a binning factor of 8, cutting to the spectral region of 1800–1000 cm^{-1}, and an extended multiplicative signal correction which uses an average of 10 background measurements of the SIRE as a reference to scale each datapoint [14, 26]. The wavenumber region of 1800-1000cm^{-1} was chosen as it contains the most spectral information. Exploratory analysis was completed using PCA followed by supervised machine learning methods including random forest (RF) and partial least squares-discriminant analysis (PLS-DA). The supervised techniques require splitting the data into training and test sets where the training set is used to identify biosignatures in a calibration phase and the model generated subsequently used for

predictions to be made on the test set [7, 12, 13, 27]. As there were no imbalances between the groups, no training set sampling adjustments were needed for classification analysis. The three groups were classified as the following: (i) GBM *versus* non-cancer, (ii) lymphoma *versus* non-cancer and (iii) GBM *versus* lymphoma.

Each classification completed using the PRFFECT toolbox had 51 reiterations to minimise standard error and to ensure a robust diagnostic model was used. The data were randomly split by patient ID at a 70/30 ratio between the training and test sets, keeping all patient spectral repeats together. The 51 reiterations shuffled the 70/30 split each time so that every patient within the whole dataset was predicted at least once.

Results

Fig 1 displays the spectral differences between the same GBM patient sample in unfiltered serum compared to each of the five molecular weight cut-off regions. Each patient serum was separated through a 100 kDa filter first, followed by 50 kDa, then 30 kDa, followed by 10 kDa and finally 3 kDa, where the filtrate (region that has passed through the filter) was analysed. This process resulted in the molecular weight regions of <100 kDa, <50 kDa, <30 kDa, <10 kDa and <3 kDa. From Fig 1, it is clear that there is a large difference within the serum spectra once the higher molecular weight components (>100 kDa) were removed. This is significant in the higher wavenumber region between 3700 cm^{-1} to 2700 cm^{-1}. However, more

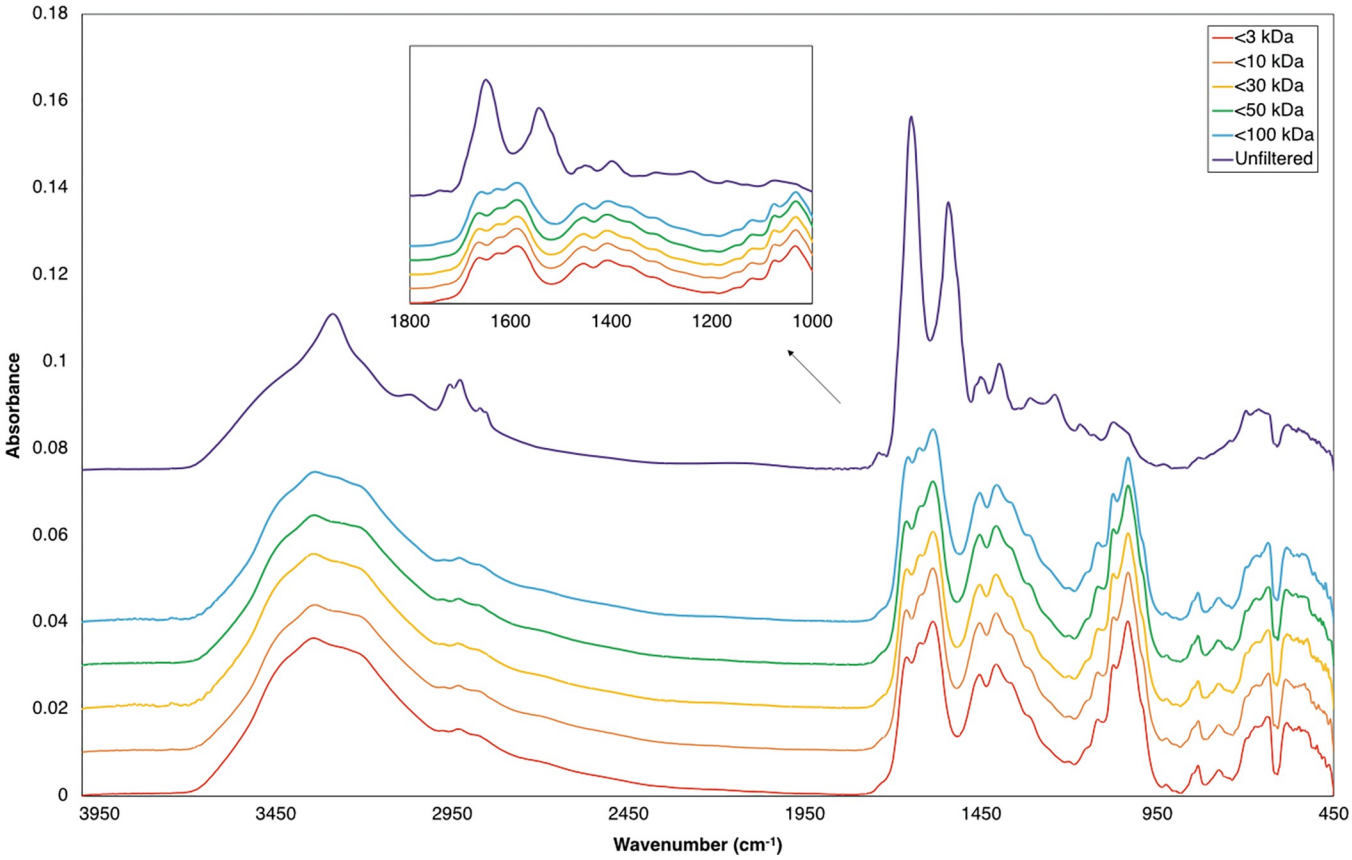

Fig 1. Example of patient serum spectra including unfiltered whole serum and each molecular weight region. Average of the 30 GBM patients shown here. The inset is the wavenumber region between 1800 cm^{-1} and 1000 cm^{-1}, which was used for all chemometrics and machine learning analyses. Spectra is offset for clearer visualisation.

https://doi.org/10.1371/journal.pone.0279669.g001

importantly there are numerous differences between 1800 cm^{-1} and 1000 cm^{-1} (Fig 1 inset), which was therefore determined as the region of interest for further analyses. The Amide I and II bands are reduced after filtration, which is perhaps unsurprising as HSA and other serum-based proteins have been removed (note HSA is ~50% of the protein content of human blood). Given this observation, as to be expected, there are no visual differences observed between the three groups of patients; GBM, lymphoma and non-cancer (S2 and S3 Figs in S1 File), however they all followed the same trend as Fig 1 once separated into specific molecular weight fractions.

The patient spectral data were subjected to both exploratory PCA and supervised classification models (RF and PLS-DA) for all three groups: (i) GBM *versus* non-cancer, (ii) lymphoma *versus* non-cancer and (iii) GBM *versus* lymphoma. Fig 2 illustrates the PCA scores results for the three groups with unfiltered serum and contains a slight separation between the classes along the second principal component (PC2). Fig 3 shows the PCA outcomes for the three groups in the molecular weight region <100 kDa, where there is no clear separation between in the classes in all groups. The PCA scores plots for <50 kDa, <30 kDa, <10 kDa and <3 kDa are contained with the S1 File and display similar results to that of the <100 kDa region, with no separation between the classes (GBM *versus* non-cancer, lymphoma *versus* non-cancer and GBM *versus* lymphoma).

Following this initial exploratory PCA, each group was analysed using the supervised learning algorithms of RF and PLS-DA. Tables 1–3 contain the sensitivity, specificity and balanced accuracy (defined as the averaged sensitivity and specificity) for each serum fraction. The PLS-DA model results are included within these tables (Tables 1–3), while the RF are provided within the S1 File.

For the GBM *versus* non-cancer there was a slight decrease in the sensitivity, specificity and balanced accuracies once the serum was filtered. However, all classification models had an overall balanced accuracy greater then 82%, suggesting that even the individual molecular weight regions of serum can predict GBM from non-cancer patients. Lymphoma *versus* non-cancer had a larger decrease in sensitivity, specificity and balanced accuracies once filtered with overall balanced accuracies ranging between 72% and 78%. When investigating between the two cancer types, GBM *versus* lymphoma, there was a significant decrease in the sensitivity, specificity and balanced accuracies. The overall balanced accuracy decreased from 91.5% to a range between 46% and 56%, suggesting the serum fractions are unreliable in being able to stratify between cancer types as there is no distinction between GBM and lymphoma.

The RF classifications for all three groups gave very similar responses to the PLS-DA. For GBM *versus* non-cancer there was more of a decrease in sensitivity, specificity and balanced accuracies once the serum was filtered. The same can be said with Lymphoma *versus* non-cancer and once again, there was no ability to stratify between the cancer types using RF model algorithms. The percentages for each groups sensitivity, specificity and balanced accuracies are displayed in the S1 File.

From these PLS-DA classification models the loadings plots were investigated in order to identify which wavenumber regions were important for the discriminations between the cohorts. Figs 4 and 5 display the PLS-DA loadings plot for each group with the unfiltered serum (Fig 4) and the first fraction of filtered serum (<100 kDa) (Fig 5). Both the first and second PLS components are shown within the figures as the majority of the spectral variance between the cohorts will be present within these two latent variables. The loadings plots for the other serum fractions (<50 kDa, <30 kDa, <10 kDa and <3 kDa) are included within the S1 File.

For the unfiltered serum the loadings plot suggests that the discrimination between GBM and non-cancer (Fig 4A) is dependent on the Amide I and Amide II proteins (region between 1700 cm^{-1} and 1500 cm^{-1}) and the glycogen/carbohydrates (1100 cm^{-1}–1000 cm^{-1}). Lymphoma

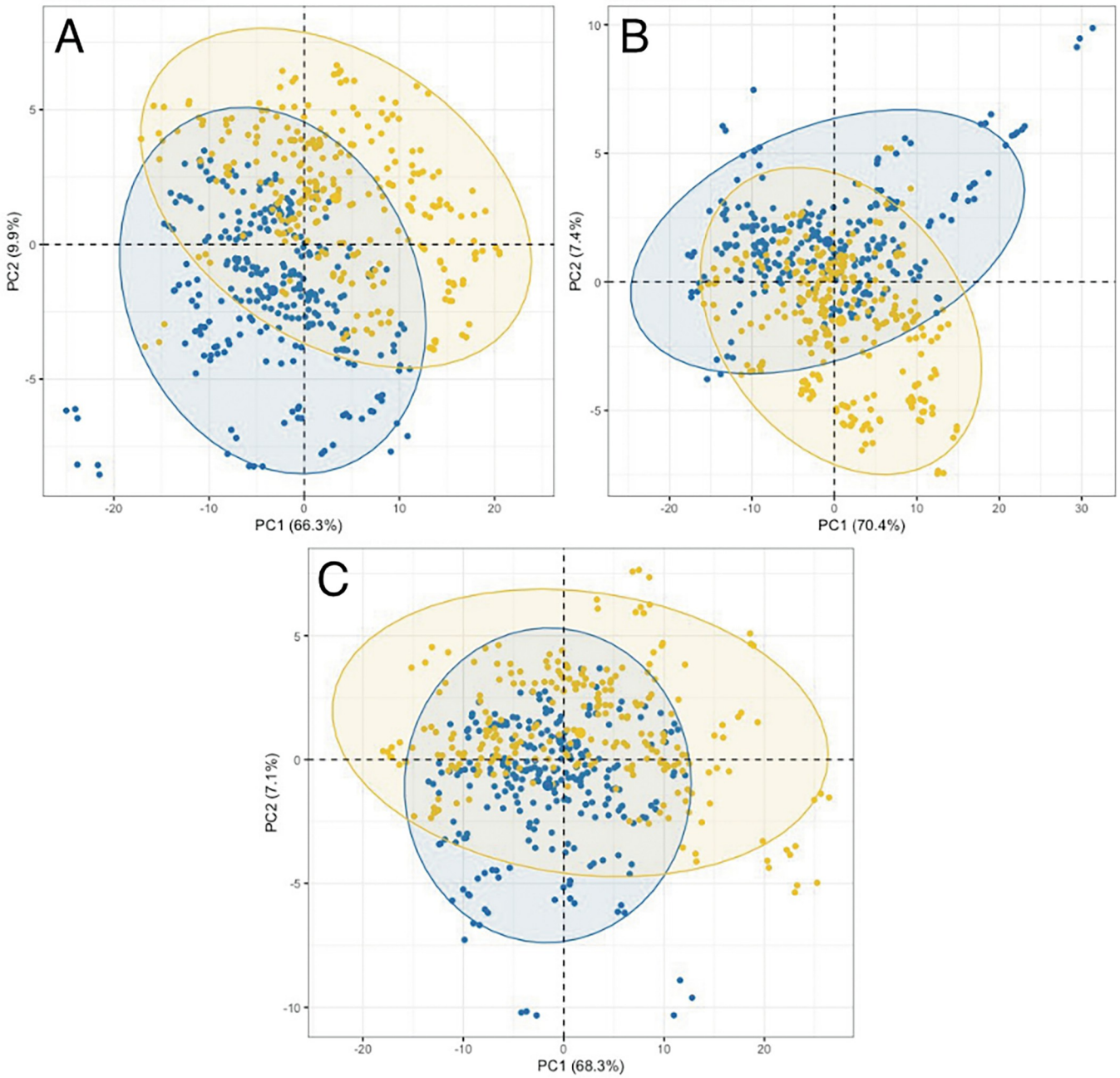

Fig 2. Principal component analysis scores plots for the unfiltered whole serum of the first and second dimensions. The three figures represent (A) GBM in blue and non-cancer in yellow, (B) lymphoma in blue and non-cancer in yellow and (C) GBM in blue and lymphoma in yellow. The eclipses in each class represent a 95% confidence interval. Values in parentheses within the axes legends are the total explained variance (TEV) for each principal component (PC).

https://doi.org/10.1371/journal.pone.0279669.g002

versus non-cancer (Fig 4B) has similar reliance on the Amide I and Amide II bands, however there is less importance in the glycogen/carbohydrates. The peak at ~1740 cm^{-1} suggests that the discrimination between lymphoma and non-cancer is also determined by the lipid components within the serum. Discriminating between cancer types, GBM *versus* lymphoma (Fig 4C), there is importance within the Amide I and Amide II region, and the glycogen/carbohydrates region. The peaks identified as important for each group is displayed in Table 4.

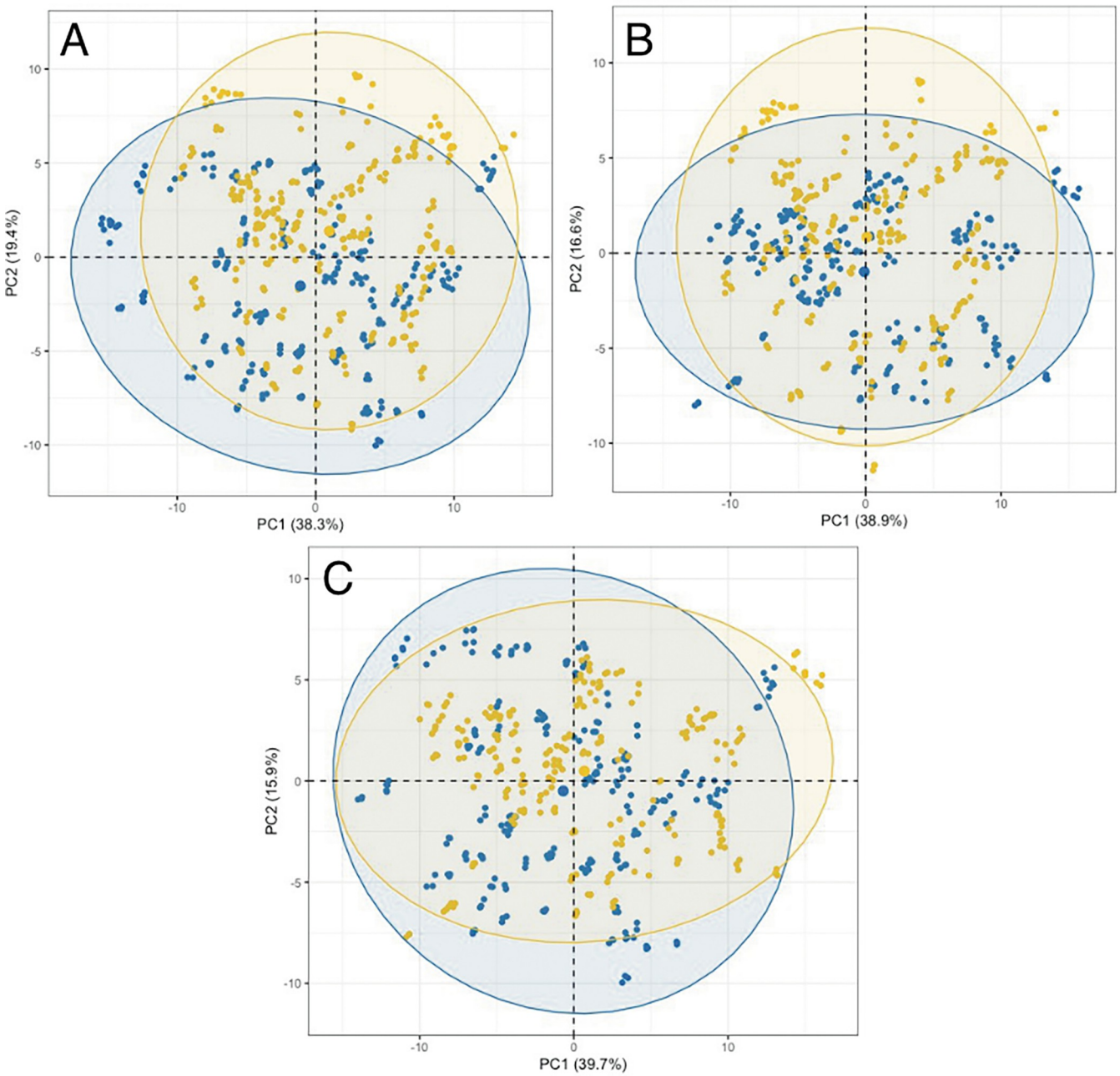

Fig 3. PCA scores plots for the filtered serum (<100 kDa) of the first and second dimensions. The three figures represent (A) GBM in blue and non-cancer in yellow, (B) Lymphoma in blue and non-cancer in yellow and (C) GBM in blue and lymphoma in yellow. The eclipses in each class represent a 95% confidence interval. Values in parentheses are the TEV for each PC.

https://doi.org/10.1371/journal.pone.0279669.g003

Once the serum was filtered the loadings plot significantly changed with what wavenumber regions were deemed important for the discrimination between patient cohorts (Fig 5). The percentage of variance from each LV will decrease with the accuracy of the model, therefore it is unreasonable to directly compare the important peaks from filtered and unfiltered serum when the whole serum will have a greater percentage of importance. It is interesting to note that more regions are deemed important within the filtered serum however the percentage of variance represented in each LV is minimal compared to the whole serum.

Table 1. Sensitivity, specificity and balanced accuracies for the PLS-DA model classification of GBM *versus* non-cancer patients. Mean, standard deviation (SD) and 95% confidence intervals (CIs) are provided.

		Sensitivity (%)			Specificity (%)			Balanced accuracy (%)		
		Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
GBM *versus* NC	Unfiltered	87.4	11.8	±3.2 84.2–90.6	92.4	8.2	±2.3 90.1–94.7	89.9	6.4	±1.8 88.1–91.7
	<100 kDa	85.5	10.7	±2.9 82.6–88.4	92.4	8.2	±2.3 90.1–94.7	89.0	6.0	±1.6 87.4–90.6
	<50 kDa	84.3	11.5	±3.2 81.1–87.5	88.9	10.4	±2.9 86.0–91.8	86.6	6.3	±1.7 84.9–88.3
	<30 kDa	84.0	13.9	±3.8 80.2–87.8	85.6	12.4	±3.4 82.2–89.0	84.8	8.7	±2.4 82.4–87.2
	<10 kDa	79.7	12.7	±3.5 76.2–83.2	85.2	13.5	±3.7 81.5–88.9	82.4	7.4	±2.0 80.4–84.4
	<3 kDa	86.8	13.7	±3.8 83.0–90.1	87.4	9.7	±2.7 84.7–90.1	87.1	8.3	±2.3 84.8–89.4

https://doi.org/10.1371/journal.pone.0279669.t001

Discussion

From the initial visual observations there was a significant difference between the spectral profiles of whole unfiltered serum and the different molecular weight fractions. This is to be expected as the initial filtration step will remove components greater than 100 kDa, including human serum albumin which comprises 50% of the protein complement of sera (30–50 g/L), and antibodies such as immunoglobulin G (IgG). IgG is one of the main components (7–16 g/L) within serum [29, 30]. The removal of serum albumin and other components within the first filtration step has a significant impact on the overall serum spectral profile. Between the 3 patient groups, GBM, lymphoma and non-cancer, visually there were no spectral differences and they all follow the same spectral trend once centrifugally filtered (Fig 1, S2 and S3 Figs in S1 File). There were visually few changes between the molecular weight fractions as most significant changes occured within the first filtration step.

Within the exploratory principal component analysis (PCA) there was slight separation between the groups along the second principal component for the unfiltered whole serum. By contrast, once filtered there was no separation between the groups of patients, demonstrated throughout all molecular weight regions. The clear distinction between groups within the unfiltered serum suggests that the higher molecular weight (>100 kDa) components within the serum play an important role for the discrimination between cancer and non-cancer or between cancer types.

Table 2. Sensitivity, specificity and balanced accuracies for the PLS-DA model classification of lymphoma *versus* non-cancer patients. Mean, standard deviation (SD) and 95% confidence intervals (CIs) are provided.

		Sensitivity (%)			Specificity (%)			Balanced accuracy (%)		
		Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
Lymphoma *versus* NC	Unfiltered	85.3	13.6	±3.7 81.6–89.0	85.8	13.3	±3.7 82.1–89.5	85.6	8.7	±2.4 83.2–88.0
	<100 kDa	66.1	18.5	±5.1 61.0–71.2	78.9	14.9	±4.1 74.8–83.0	72.5	11.5	±3.2 69.3–75.7
	<50 kDa	66.9	19.7	±5.4 61.5–72.3	83.7	12.8	±3.5 80.2–87.2	75.3	10.9	±3.0 72.3–78.3
	<30 kDa	74.2	17.4	±4.7 69.4–79.0	80.6	13.8	±3.8 76.8–84.4	77.4	10.4	±2.9 74.5–80.3
	<10 kDa	70.6	16.4	±4.5 66.1–75.1	81.5	12.1	±3.3 78.2–84.8	76.0	10.8	±3.0 76.0–79.0
	<3 kDa	65.7	15.2	±4.2 61.5–69.9	82.8	13.6	±3.7 79.1–86.5	74.2	9.6	±2.6 71.6–76.8

https://doi.org/10.1371/journal.pone.0279669.t002

Table 3. Sensitivity, specificity and balanced accuracies for the PLS-DA model classification of GBM *versus* Lymphoma patients. Mean, standard deviation (SD) and 95% confidence intervals (CIs) are provided.

		Sensitivity (%)			Specificity (%)			Balanced accuracy (%)		
		Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
GBM *versus* lymphoma	Unfiltered	97.1	5.4	±1.5 95.6–98.6	86.0	10.5	±2.9 83.1–88.9	91.5	6.2	±1.7 89.8–93.2
	<100 kDa	52.9	17.7	±4.9 48.0–57.8	53.2	17.7	±4.9 48.3–58.1	53.1	9.7	±2.7 50.4–55.8
	<50 kDa	53.8	18.9	±5.2 48.6–59.0	56.0	16.6	±4.6 51.4–60.6	54.9	11.7	±3.2 51.7–58.1
	<30 kDa	54.4	18.8	±5.2 49.2–59.6	57.1	21.0	±5.8 51.3–62.9	55.8	12.1	±3.3 52.5–59.1
	<10 kDa	55.4	17.9	±4.9 50.5–60.3	44.5	19.7	±5.4 39.1–49.9	50.0	10.9	±3.0 47.0–53.0
	<3 kDa	27.5	15.0	±4.1 23.4–31.6	65.5	20.4	±5.6 59.9–71.1	46.5	11.0	±3.0 43.5–49.5

https://doi.org/10.1371/journal.pone.0279669.t003

These observations can be confirmed through the supervised classification analysis where each group of patients was stratified using both PLS-DA and RF machine learning algorithms. For the unfiltered serum the classifications between the two patient groups (GBM *versus* non-cancer, lymphoma *versus* non-cancer or GBM *versus* lymphoma) all gave overall balanced accuracies above 85%. When focusing on GBM *versus* non-cancer the sensitivities, specificities and balanced accuracies of the models remained around or greater then 80%; however, none of the molecular weight filtrates gave percentages as high as the unfiltered whole serum. Lymphoma *versus* non-cancer had a more noticeable decrease in sensitivity, specificity and balanced accuracies once the patient serum was filtered. These

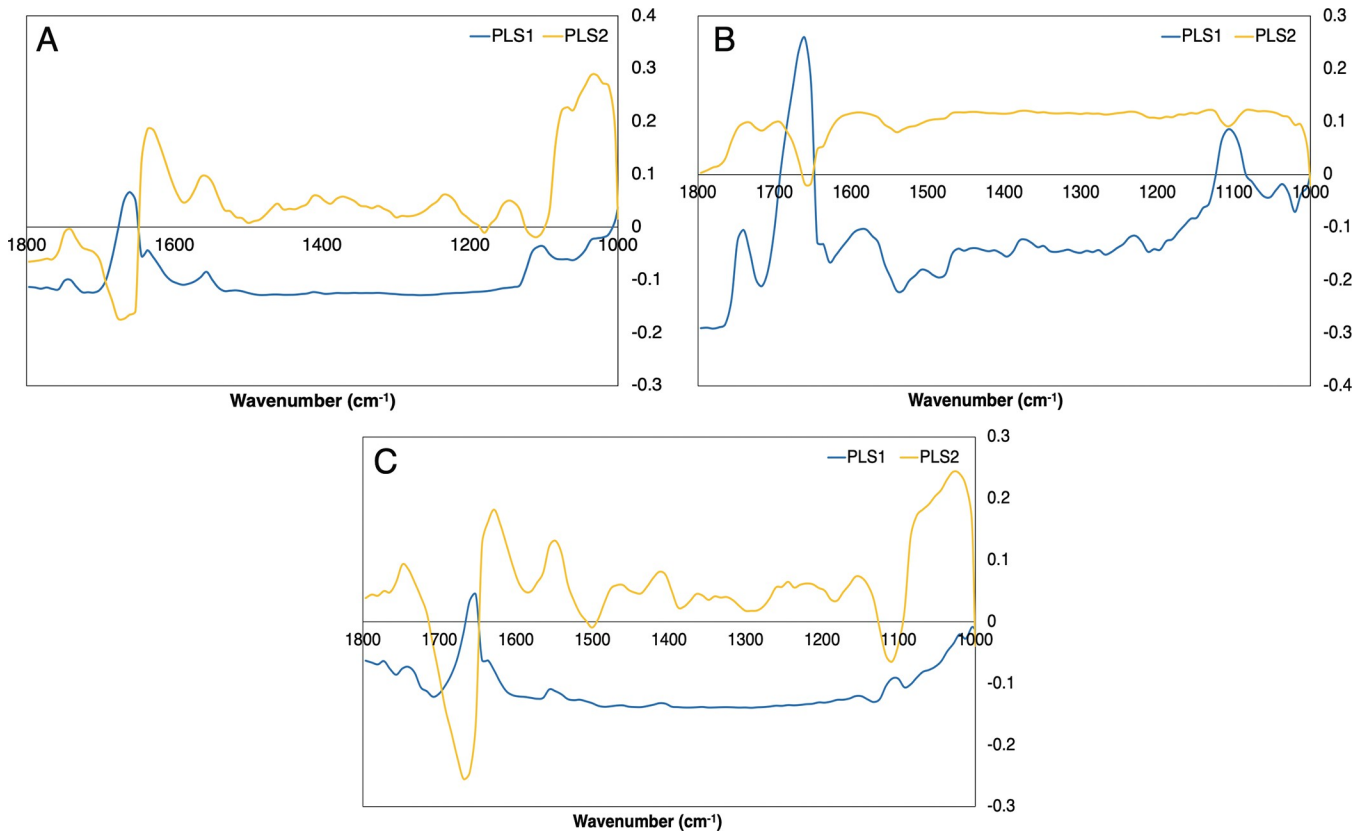

Fig 4. PLS loadings plots for the 1st and 2nd latent variables (LVs) for the unfiltered whole serum. (A) GBM *versus* non-cancer, (B) Lymphoma *versus* non-cancer and (C) GBM *versus* lymphoma.

https://doi.org/10.1371/journal.pone.0279669.g004

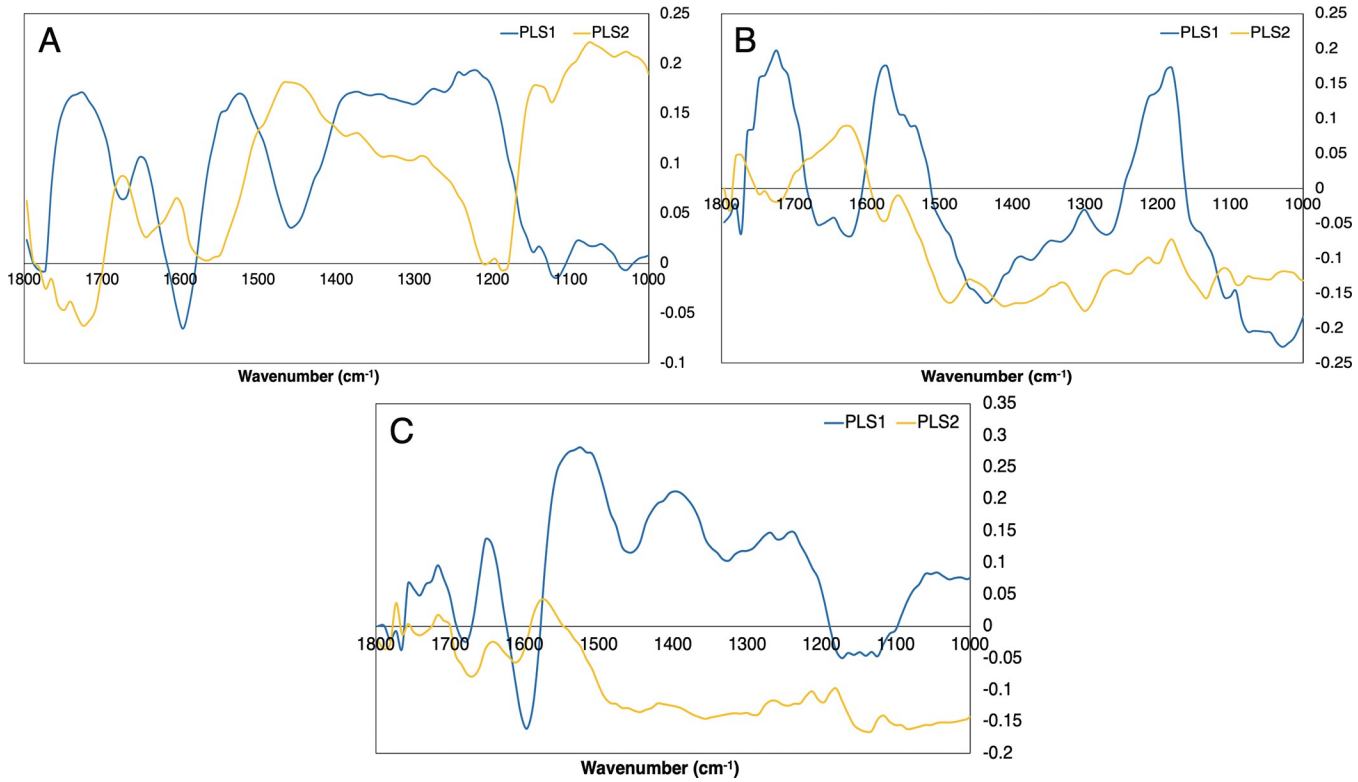

Fig 5. PLS loadings plot for the 1st and 2nd LVs for the filtered serum (<100 kDa). (A) GBM *versus* non-cancer, (B) Lymphoma *versus* non-cancer and (C) GBM *versus* lymphoma.

https://doi.org/10.1371/journal.pone.0279669.g005

Table 4. Top wavenumbers for each group in unfiltered serum classifications. Tentative biochemical assignments and their corresponding vibrational modes are included [28].

	Wavenumber (cm^{-1})	Tentative assignment	Vibrational modes
GBM *versus* NC	1668.5	Amide I of proteins	ν(C = O), ν(C-N), δ(N-H)
	1660.5	Amide I of proteins	ν(C = O), ν(C-N), δ(N-H)
	1628.5	Amide I of proteins	ν(C = O), ν(C-N), δ(N-H)
	1556.5	Amide II of proteins	δ(N-H), ν(C-N), δ(C-O), ν(C-C)
	1100.5	Nucleic acids	$\nu(PO^{2-})$
	1028.5	Glycogen	ν(C-O), ν(C-C), def(C-OH)
Lymphoma *versus* NC	1740.5	Lipids	ν(C-O)
	1660.5	Amide I of proteins	ν(C = O), ν(C-N), δ(N-H)
	1652.5	Amide I of proteins	ν(C = O), ν(C-N), δ(N-H)
	1580.5	Amide II of proteins	δ(N-H), ν(C-N), δ(C-O), ν(C-C)
	1108.5	Carbohydrate	ν(C-O), ν(C-C)
GBM *versus* Lymphoma	1668.5	Amide I of proteins	ν(C = O), ν(C-N), δ(N-H)
	1660.5	Amide I of proteins	ν(C = O), ν(C-N), δ(N-H)
	1628.5	Amide I of proteins	ν(C = O), ν(C-N), δ(N-H)
	1548.5	Amide II of proteins	δ(N-H), ν(C-N), δ(C-O), ν(C-C)
	1108.5	Carbohydrate	ν(C-O), ν(C-C)
	1020.5	Glycogen	ν(C-O), ν(C-C)

ν = stretching

δ = bending; def = deformation

https://doi.org/10.1371/journal.pone.0279669.t004

remained at 65% and greater, however, as with the GBM *versus* non-cancer cohort the unfiltered serum outperformed each filtrate. Between the two cancer types, GBM versus lymphoma, there was a significant decrease in stratification ability once the serum was filtered. The overall balanced accuracies of 50% suggest that there are no discriminatory features within the serum to identify between a GBM or lymphoma brain cancer patient once the components above 100 kDa were removed.

For all classifications the unfiltered whole serum performed the greatest which suggests that the higher molecular weight components are needed for discriminatory ability between these binary cohorts. From a clinical application point of view, this is preferable as the extra pre-analytical steps to include the filtration is more time consuming and harder to translate into a clinic ready test.

Supporting information

S1 File. Contains all the supporting figures.
(DOCX)

Acknowledgments

The authors would like to thank both the Walton Centre NHS Foundation Trust and the Lancashire Teaching Hospitals NHS Trust for the collaboration and access to patient samples.

Author Contributions

Conceptualization: Ashton G. Theakstone, Paul M. Brennan, Matthew J. Baker.

Data curation: Ashton G. Theakstone.

Formal analysis: Ashton G. Theakstone.

Funding acquisition: Matthew J. Baker.

Investigation: Ashton G. Theakstone.

Methodology: Ashton G. Theakstone, Matthew J. Baker.

Project administration: Matthew J. Baker.

Resources: Ashton G. Theakstone, Michael D. Jenkinson.

Software: Royston Goodacre.

Supervision: Matthew J. Baker.

Validation: Ashton G. Theakstone.

Visualization: Ashton G. Theakstone.

Writing – original draft: Ashton G. Theakstone.

Writing – review & editing: Ashton G. Theakstone, Paul M. Brennan, Michael D. Jenkinson, Royston Goodacre, Matthew J. Baker.

References

1. Ozawa M, Brennan PM, Zienius K, Kurian KM, Hollingworth W, Weller D, et al. The usefulness of symptoms alone or combined for general practitioners in considering the diagnosis of a brain tumour: a case-control study using the clinical practice research database (CPRD) (2000–2014). BMJ Open. 2019; 9 (8):e029686. Epub 2019/09/01. https://doi.org/10.1136/bmjopen-2019-029686 PMID: 31471440; PubMed Central PMCID: PMC6720478.

2. Hamilton W, Kernick D. Clinical features of primary brain tumours: a case-control study using electronic primary care records. Br J Gen Pract. 2007; 57(542):695–9. Epub 2007/09/01. PMID: 17761056; PubMed Central PMCID: PMC2151783.
3. Latinovic R. Headache and migraine in primary care: consultation, prescription, and referral rates in a large population. Journal of Neurology, Neurosurgery & Psychiatry. 2005; 77(3):385–7. https://doi.org/10.1136/jnnp.2005.073221 PMID: 16484650
4. Swann R, McPhail S, Witt J, Shand B, Abel GA, Hiom S, et al. Diagnosing cancer in primary care: results from the National Cancer Diagnosis Audit. Br J Gen Pract. 2018; 68(666):e63–e72. https://doi.org/10.3399/bjgp17X694169 PMID: 29255111
5. Cameron JM, Brennan PM, Antoniou G, Butler HJ, Christie L, Conn JJA, et al. Clinical validation of a spectroscopic liquid biopsy for earlier detection of brain cancer. Neuro-Oncology Advances. 2022; 4(1): vdac024. https://doi.org/10.1093/noajnl/vdac024 PMID: 35316978
6. Brennan PM, Butler HJ, Christie L, Hegarty MG, Jenkinson MD, Keerie C, et al. Early diagnosis of brain tumours using a novel spectroscopic liquid biopsy. Brain Commun. 2021; 3(2). https://doi.org/10.1093/braincomms/fcab056 PMID: 33997782
7. Smith BR, Ashton KM, Brodbelt A, Dawson T, Jenkinson MD, Hunt NT, et al. Combining random forest and 2D correlation analysis to identify serum spectral signatures for neuro-oncology. Analyst. 2016; 141:3668–78. https://doi.org/10.1039/c5an02452h PMID: 26818218
8. Baker MJ, Trevisan J, Bassan P, Bhargava R, Butler HJ, Dorling KM, et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc. 2014; 9(8):1771–91. https://doi.org/10.1038/nprot.2014.110 PMID: 24992094
9. Hands JR, Abel P, Ashton K, Dawson T, Davis C, Lea RW, et al. Investigating the rapid diagnosis of gliomas from serum samples using infrared spectroscopy and cytokine and angiogenesis factors. Analytical and Bioanalytical Chemistry. 2013; 405(23):7347–55. https://doi.org/10.1007/s00216-013-7163-z PMID: 23831829
10. Hands JR, Dorling KM, Abel P, Ashton KM, Brodbelt A, Davis C, et al. Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples. J Biophotonics. 2014; 7(3–4):189–99. https://doi.org/10.1002/jbio.201300149 PMID: 24395599
11. Hands JR, Clemens G, Stables R, Ashton K, Brodbelt A, Davis C, et al. Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy. J Neurooncol. 2016; 127(3):463–72. Epub 2016/02/15. https://doi.org/10.1007/s11060-016-2060-x PMID: 26874961; PubMed Central PMCID: PMC4835510.
12. Cameron JM, Butler HJ, Smith BR, Hegarty MG, Jenkinson MD, Syed K, et al. Developing infrared spectroscopic detection for stratifying brain tumour patients: glioblastoma multiforme vs. lymphoma. Analyst. 2019; 144(22):6736–50. https://doi.org/10.1039/c9an01731c PMID: 31612875
13. Butler HJ, Brennan PM, Cameron JM, Finlayson D, Hegarty MG, Jenkinson MD, et al. Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. Nat Commun. 2019; 10(1):1–9. https://doi.org/10.1038/s41467-019-12527-5 PMID: 31594931
14. Theakstone AG, Brennan PM, Jenkinson MD, Mills SJ, Syed K, Rinaldi C, et al. Rapid Spectroscopic Liquid Biopsy for the Universal Detection of Brain Tumours. Cancers. 2021; 13(15):3851. https://doi.org/10.3390/cancers13153851 PMID: 34359751
15. Group PH. Harmonisation of Reference Intervals. Clinical Biochemistry Outcomes 2011.
16. Hu S, Loo JA, Wong DT. Human body fluid proteome analysis. PROTEOMICS. 2006; 6(23):6326–53. https://doi.org/10.1002/pmic.200600284 PMID: 17083142
17. Adkins JN, Varnum SM, Auberry KJ, Moore RJ, Angell NH, Smith RD, et al. Toward a Human Blood Serum Proteome. Molecular & Cellular Proteomics. 2002; 1(12):947–55. https://doi.org/10.1074/mcp.m200066-mcp200 PMID: 12543931
18. Bonnier F, Baker MJ, Byrne HJ. Vibrational spectroscopic analysis of body fluids: avoiding molecular contamination using centrifugal filtration. Analytical Methods. 2014; 6(14):5155. https://doi.org/10.1039/c4ay00891j
19. Bonnier F, Blasco H, Wasselet C, Brachet G, Respaud R, Carvalho LFCS, et al. Ultra-filtration of human serum for improved quantitative analysis of low molecular weight biomarkers using ATR-IR spectroscopy. Analyst. 2017; 142:1285–98. https://doi.org/10.1039/c6an01888b PMID: 28067340
20. Butler HJ, Smith BR, Fritzsch R, Radhakrishnan P, Palmer DS, Baker MJ. Optimised spectral pre-processing for discrimination of biofluids via ATR-FTIR spectroscopy. Analyst. 2018; 143(24):6121–34. https://doi.org/10.1039/c8an01384e PMID: 30484797
21. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32. https://doi.org/10.1023/a:1010933404324

22. Ali J, Khan R, Ahmad N, Maqsood I. Random Forests and Decision Trees. Int J Comput Sci Issues. 2012; 9:272–8.

23. Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods. 2013; 5(16):3790. https://doi.org/10.1039/c3ay40582f

24. Lee LC, Liong C-Y, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst. 2018; 143(15):3526–39. https://doi.org/10.1039/c8an00599k PMID: 29947623

25. Smith BR, Baker MJ, Palmer DS. PRFFECT: A versatile tool for spectroscopists. Chemom Intell Lab Syst. 2018; 172:33–42. https://doi.org/10.1016/j.chemolab.2017.10.024

26. Theakstone AG, Brennan PM, Ashton K, Czeiter E, Jenkinson MD, Syed K, et al. Vibrational Spectroscopy for the Triage of Traumatic Brain Injury Computed Tomography Priority and Hospital Admissions. J Neurotrauma. 2022; 39(11–12):773–83. Epub 2022/03/04. https://doi.org/10.1089/neu.2021.0410 PMID: 35236121; PubMed Central PMCID: PMC9225408.

27. Sala A, Anderson DJ, Brennan PM, Butler HJ, Cameron JM, Jenkinson MD, et al. Biofluid diagnostics by FTIR spectroscopy: A platform technology for cancer detection. Cancer Lett. 2020; 477:122–30. https://doi.org/10.1016/j.canlet.2020.02.020 PMID: 32112901

28. Movasaghi Z, Rehman S, Ur Rehman DI. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. Applied Spectroscopy Reviews. 2008; 43(2):134–79. https://doi.org/10.1080/05704920701829043

29. Leeman M, Choi J, Hansson S, Storm MU, Nilsson L. Proteins and antibodies in serum, plasma, and whole blood—size characterization using asymmetrical flow field-flow fractionation (AF4). Analytical and Bioanalytical Chemistry. 2018; 410(20):4867–73. https://doi.org/10.1007/s00216-018-1127-2 PMID: 29808297

30. Gonzalez-Quintela A, Alende R, Gude F, Campos J, Rey J, Meijide LM, et al. Serum levels of immunoglobulins (IgG, IgA, IgM) in a general adult population and their relationship with alcohol consumption, smoking and common metabolic abnormalities. Clin Exp Immunol. 2008; 151(1):42–50. Epub 2007/11/17. https://doi.org/10.1111/j.1365-2249.2007.03545.x PMID: 18005364; PubMed Central PMCID: PMC2276914.