MQP-BIO-DSA-5242

OPTIMIZATION OF PROTEOMIC ANALYSIS METHODS OF COMPLEX BIOLOGICAL MIXTURES

A Major Qualifying Project Report

Submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

in

Biology and Biotechnology

by

Emma-Jane Turton

eCDR Deadline: April 27, 2017

APPROVED:

Kate Zhang, PhD
Department of Biomarkers and Bioanalytics
Sanofi Genzyme
MAJOR ADVISOR

David Adams, PhD
Biology and Biotechnology
WPI Project Advisor

ABSTRACT

The concept of proteomics has played a crucial role in understanding biological processes. Biomarker discovery is important for disease diagnosis and treatment, monitoring. This project evaluated the effect of protein digest conditions and sample contamination on biomarker quantitation in biological mixtures. The biological mixtures tested included cerebrospinal fluid (CSF) and plasma. Tandem mass spectrometry was used to analyze and identify specific marker peptides. Based on the data collected, the optimal method for peptide quantitation is dependent on the matrix tested, and varies between unbiased and targeted approaches. In CSF, the best protein digestion conditions included the use of Rapigest surfactant reagent, while plasma digestion was more compatible with 2,2,2-trifluoroethanol solvent during the digestion. The CSF sample spiked with trace blood level was used as a positive control to monitor blood contamination during CSF collection and process. Blood contamination in CSF was shown to not be a major concern in the current targeted protein analysis since only one of 56 targeted peptides was slighted affected.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would first like to thank Pavlina Wolf of Sanofi Genzyme for working diligently to set up this project, and for all of the skills and knowledge she has passed on to me throughout this entire experience. Also, thank you to Melissa Rotunno for her efforts and assistance in collaboration with this project and sharing her expertise in targeted proteomics. Thank you to Petra Oliva, Monica Lane and the rest of the Biomarkers and Bioanalytics team for all of your help and encouragement over the past year. Additionally, thank you to Kate Zhang, who provided the opportunity to work in her department. Finally, I would like to thank Professor David Adams for not only this project, but all of his guidance throughout my four years at WPI.

BACKGROUND

Proteomics

Proteins are macromolecules present in all living things (Chandrasekhar et al., 2014). They are composed of amino acids and can be identified by their sequence, structure and function (Chandrasekhar et al., 2014). The study of proteins at a large scale and their properties is termed proteomics, and is becoming more valuable in scientific research (Chandrasekhar et al., 2014). Information found in a person's genome and proteome can potentially be used to identify specific genes and proteins that are associated with disease (Chandrasekhar et al., 2014). A proteome is the entire complement of proteins that is being expressed, or could be expressed, by a cell, tissue, or organism. A proteome is complex; in fact, it can consist of about 100,000 proteins (Kellner, 2000). This complexity can lead to high variability from human to human, and generate unlimited phenotypes (Kellner, 2000). Internal and external forces such as environmental stressors, disease or drugs can affect proteome composition and individual protein expression. The changes in protein abundances due to these effects, as well as the complete presence or absence of a particular protein can be measured and provide important clues of the health of the organism (Kellner, 2000). Proteomics can be more valuable in understanding an organism than genomics because it seeks to explain which proteins interact under dynamic conditions (Chandrasekhar et al., 2014). While protein expression can change over time due to various conditions, an individual genome remains fairly constant, explaining why "proteomics is often considered as the advanced step in the study of biological systems," (Chandrasekhar et al., 2014).

Proteomics can be divided into three types of study: expression proteomics, structural proteomics, and functional proteomics (Chandrasekhar et al., 2014). Expression proteomics involves comparing protein expression and patterns in normal and diseased samples (Chandrasekhar et al., 2014). In this type of analysis, abnormal cells are tested to determine which proteins are over-expressed and under-expressed to potentially identify biomarkers and therapeutic targets (Chandrasekhar et al., 2014). Common methods of quantitation for expression proteomics include 2-D gel electrophoresis and mass spectrometry (Chandrasekhar et al., 2014). Structural proteomics is used to predict structural properties and three dimensional shapes of proteins (Chandrasekhar et al., 2014). Using technologies like X-ray crystallography and NMR spectroscopy, it is possible to identify proteins present in membranes, ribosomes and other cellular organelles (Chandrasekhar et al., 2014). Finally, functional proteomics aims to identify protein interactions and function (Chandrasekhar et al., 2014). Determining how a specific "unknown" protein functions with respect to a known protein complex and mechanism could be indicative of its own biological function (Chandrasekhar et al., 2014).

Matrices

i. CSF

CSF is a clear bodily fluid surrounding parts of the brain and spinal cord. CSF filters into the central nervous system (CNS) across the choroid plexus located at the brain ventricles. The fluid circulates around the brain and spinal cord, and is reabsorbed back into the blood stream via the arachnoid granulations and other

avenues. Full CSF turnover in a human takes about 6-8 hours. The fluid is mostly water, ions, and glucose, but also contains 0.2 – 0.8 mg/mL of proteins. CSF collected near the choroid plexus has been shown to contain proteins that originate in the blood, while CSF collected from the brain tissue is enriched in central nervous system proteins (Zhang et al., 2015). Lumbar puncture can provide a cross sample of these proteins for analysis.

Scientists believe that alterations in the amount of proteins found in CSF could potentially lead to discoveries of biomarkers for neurological disorders (Zhang et al., 2015). For example, in a 2009 study for protein biomarkers of Alzheimer's disease, Tapiola et al. concluded that a combination of extremely low amyloid beta 42 and heightened tau protein levels in cerebrospinal fluid is to be considered a biomarker for Alzheimer's disease associated pathologic changes in the brain (Tapiola et al., 2009). To begin a study of potential biomarkers for neurological diseases, it is important to know what a normal CSF proteome looks like. In 2002, Albert Sickmann et al., released a detailed overview of normal CSF proteins in which they used MALDI mass fingerprinting to find 85 total proteins (Sickmann et al., 2002). Since then, researchers have been able to identify over 3200 unique proteins using high resolution mass spectrometry (Zhang et al., 2015). It should be noted that a few high abundance proteins tend to make up most of the total protein amount (Zhang et al., 2015). In order to properly utilize CSF as a matrix to test for neurological biomarkers, either the high abundance proteins need to be depleted, or the sample needs to be enriched in the lower abundance proteins for best quantitation (Zhang et al., 2015).

ii. Plasma

Plasma is the highly proteinated liquid portion of blood that remains after the removal of cellular components. Plasma is responsible for about 55% of total blood volume (American Red Cross, 2017). This colorless and viscous fluid functions in many ways, including blood pressure regulation and immune responses (American Red Cross, 2017). Blood is readily available because it can be found almost anywhere in the body and is constantly being produced (Dayon and Kussman, 2013). Obtaining blood is a much less invasive process compared to cerebrospinal fluid, making it an ideal matrix to be measured in routine clinical practice and research (Dayon and Kussman, 2013). Blood plasma is a universal indicator of specific sub-proteomes in the body, but proteomic analysis presents some unique challenges as discussed previously, including the difficulty quantifying low abundance proteins in a biological fluid containing a few high abundance proteins (Jacobs et al., 2005).

In a typical disease biomarker study, a list of biomarker candidates is created by comparing disease (case) and control samples using unbiased proteomic approaches (Dayon and Kussman, 2013). This is then followed by a targeted approach to refine the potential markers identified through unbiased approach (Dayon and Kussman, 2013). These types of mass spectrometry analyses can be challenging due to the complexity and range of the plasma proteome (Dayon and Kussman, 2013). Even more than in other biofluids, the highly abundant proteins found in blood tend to dominate MS analysis (Dayon and Kussman, 2013) which can prevent low abundance peptides from being accurately quantified or potential biomarkers from being discovered (Dayon and Kussman, 2013).

Mass Spectrometry

Mass spectrometry (MS) is a highly sensitive analytical technique that has become widely used across many fields including proteomics and genomics (Thermo Fisher, 2017). In fact, much of what has come to be known about the proteome has happened within the last two decades with considerable thanks to mass spectrometry development (Thermo Fisher, 2017). Mass spectrometry identifies molecules based on their mass-to-charge ratio (m/z), and was originally used by researchers to measure isotope abundance in biological mixtures (Thermo Fisher, 2017). Now, the technique can be used in a variety of ways, such as peptide sequencing and the structural analysis of nucleotides (Thermo Fisher, 2017).

Every mass spectrometer has an ion source, mass analyzer and ion detector (Thermo Fisher, 2017). When a sample is introduced into the ion source of the MS, the molecule is ionized and propelled across an electric field produced by the mass analyzer (Thermo Fisher, 2017). The ions deflect the charge in a path that is based on their m/z ratio, producing an output for analysis (Thermo Fisher, 2017). Tandem mass spectrometry (MS/MS) (**Figure-1**) can be used for protein and peptide quantitation (Thermo Fisher, 2017). The first segment of MS (MS1) generates a precursor ion spectrum scanning through all ionized peptides within the sample (Thermo Fisher, 2017). The MS1 generated ions are then fragmented by collision-induced fragmentation (CID) in the second segment of the mass analyzer where the signature fragment ion of the molecule can be detected (Thermo Fisher, 2017). These fragmented MS2 spectra create a

confirmation for the peptides analyzed in MS1 to give a more concrete identification of peptides (Thermo Fisher, 2017).

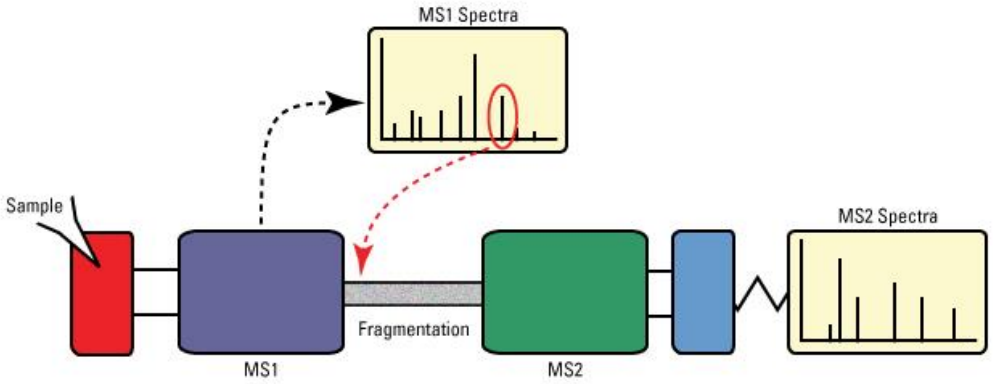


Figure 1: Diagram of Tandem Mass Spectrometry. A sample is applied (red, diagram left), MS1 (purple) creates precursor ion spectra (upper diagram), ions undergo collision induced fragmentation (gray), and MS2 (green) generates fragmented ion spectrum (diagram right) for data processing (Thermo Fisher, 2017).

i. Unbiased vs. Targeted Mass Spectrometry

Mass spectrometry-based proteomics can be classified into two methods: unbiased analysis for broad discovery and targeted quantification of selected protein biomarkers (**Figure-2**) (Doerr, A 2013). The discovery-based strategy (diagram, right side) typically analyses large numbers of proteins to cover protein in high and medium level of concentration. For example, large numbers of proteins could be analyzed in two populations such as tumor versus normal surrounding tissue to discern potential differences. Due to the high number of proteins involved, depletion of highly abundant proteins is an effective step to enrich relatively lower abundant protein therefore broadening protein coverage (Thermo Fisher, 2017). Protein mixtures are first enzymatically digested to generate peptides in the sample preparation (Thermo Fisher,

2017). There, peptide mixtures are introduced into a mass spectrometer and fragmented inside the mass analyzer to generate many series of fragment ions along the protein sequence (Thermo Fisher, 2017). The spectra are then matched to a predetermined library to identify the corresponding peptide sequences (Doerr, A. 2013). These sequences are then matched to the proteins from which they originated (Doerr, A. 2013).

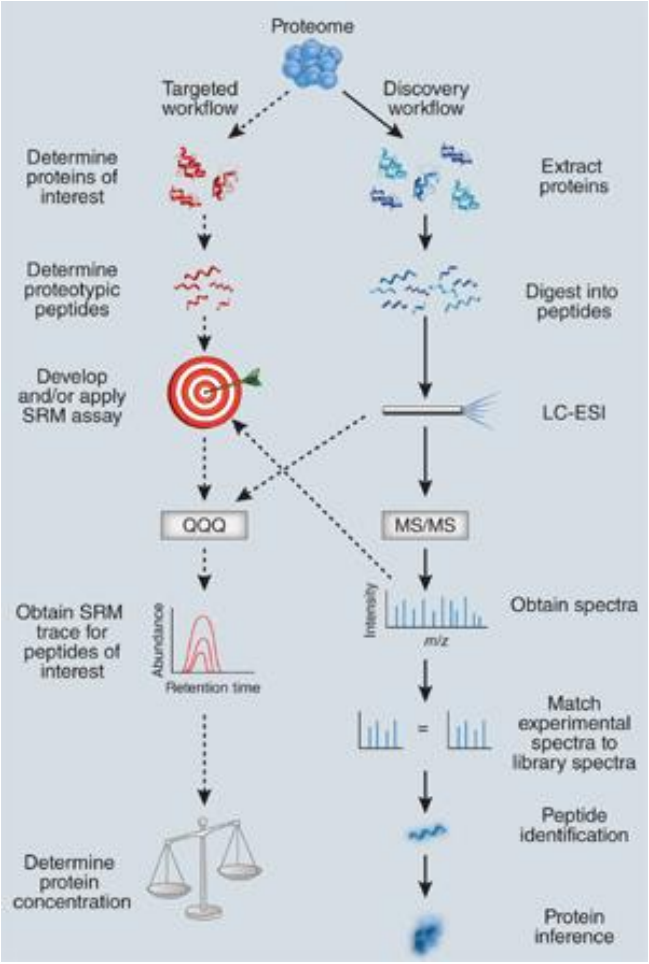


Figure 2: Comparison of Targeted versus Discovery Mass Spectrometry-Based Proteomics. The targeted approach is shown on the left, and the discovery approach on the right. LC-ESI (liquid chromatography-electrospray ionization) MS/MS (tandem mass spectrometry) (Doerr, 2013).

In the targeted proteomics approach (diagram left side), the goal is to "target" only a set of specific proteins with greater sensitivity, accuracy, reproducibility, and throughput (Doerr, 2013). In this case, the mass spectrometer is essentially told which peptide ions are to be detected, and therefore, the assays must be developed on a protein-to-protein basis (Doerr, 2013). Targeted mass spectrometry techniques for proteomic analysis have become increasingly favorable in quantifying elements of biological samples in pharmaceutical and diagnostic environments (Thermo Fisher, 2017). Although these two approaches may call for different procedures and instruments, they are often used together to validate protein identification (Thermo Fisher, 2017). "Targeted proteomics often follows discovery proteomics to quantitate or validate specific proteins found during discovery screening," (Thermo Fisher, 2017).

Detection of Low Abundance Proteins

As mentioned previously, during MS analyses low abundance proteins (LAPs) often are overshadowed by a few higher abundance proteins, especially in plasma where albumin, for instance, constitutes about 55% of total protein (Anderson, 2002). This abundance can cause difficulty in identifying potential disease biomarkers, as the high abundance proteins obscure the low abundance proteins. In fact, the plasma proteome is so complex that its dynamic range extends across >10 orders of magnitude in concentration (Anderson, 2002). To put this into perspective, the normal concentration range for the highly abundant protein serum albumin is 35-50 mg/ml (35-50 x 10^9 pg/ml) (Anderson, 2002), while the normal range of interleukin 6 (a clinically useful low abundance protein) is around 0-5 pg/ml (Anderson, 2002). The two proteins differ in

plasma abundance by about 10^{10}, however are both extremely meaningful in identifying diseases (Anderson, 2002).

In order to successfully analyze LAPs, researchers have to use either advanced chromatographic and mass spectrometric methods or additional sample processing. Mass spectrometry and tandem mass spectrometry have proved beneficial as a way to help solve this issue (Anderson, 2002). Gygi et al. introduced multi-dimensional chromatography (cation exchange, biotin affinity, and reverse phase) as well as isotope-coded affinity tags to label low abundance peptides in complex biological mixtures (Gygi et al., 2002). Another approach by Keshishian et al. involved using multiple reaction monitoring with stable isotope dilution mass spectrometry (MRM/SID-MS) to identify LAPs without peptide enrichment (Keshishian et al, 2007).

During sample processing, high abundance proteins can be removed using immuno-depletion (Millioni et al., 2011). Ultimately, the goal is to deplete as many high abundance proteins as possible without also losing proteins of interest in the process (Echan, 2005). Various depletion methods have been developed, including a dye-based (Cibacron blue) stain for binding and depleting albumin, and Protein A or G for removing immunoglobulins (Echan, 2005). Several different protein depletion kits are commercially available today. Most are designed with compatibility for plasma and MS analyses (Filip, 2015).

PROJECT PURPOSE

Studying the human proteome can play a vital role in identifying potential biological markers for a variety of diseases. In doing so, it is important to develop research methods that can identify and quantify the maximum amount of proteins within the matrices of interest (e.g. plasma or CSF). Biomarkers can be either highly abundant proteins or proteins with low abundance. The purpose of this project was to optimize analytical methods that allow for the accurate detection of low abundance proteins in small volumes of body fluids, and to maximize the number of proteins identified. As collecting CSF can be an invasive procedure in which blood contamination is common, this project was also designed to measure the effect of various levels of blood contamination in CSF on biomarker levels.

METHODS

Sample Preparation

CSF QC Pool Preparation: Human CSF from 15 individuals (Precision Med) was pooled, treated 1:100 with protease and phosphatase inhibitors (Thermo Scientific), aliquoted for single use, and stored at -80ºC until digest.

BCA Assay: Total protein concentration of each pooled CSF sample was determined using the Pierce BCA protein assay kit against seven standards (ranging from 125-2000 µg/mL) of pre-diluted BSA mixtures. Ten microliters of each standard or sample was added to a microplate well, 200 µl of working reagent was then added, the plate was sealed, shaken for 30 seconds, and then incubated for 30 minutes at 37ºC. The plate was cooled to room temperature and absorbance was measured at 562 nm.

Proteomic Digest Method Optimization: Twenty micrograms of total protein for each sample was digested. The samples were incubated in acetone for 1 hr at -20ºC to precipitate the protein, which was pelleted by spinning at 20,000 g at 4ºC for 20 minutes. The supernatant was aspirated, and samples were washed with acetone three times. After the last wash, samples were dried down using a SpeedVac until the acetone completely evaporated. Samples were then split into two groups of six. The first group underwent one freeze-thaw cycle before digestion and the second group was digested without freeze thaw.

Each group of six samples (with and without a freeze-thaw step) was then split in half, and digested in one of two detergents 2,2,2-triflouroethanol (TFE) (Sigma) or Rapigest (Waters). Therefore, three frozen CSF samples and three fresh samples were reconstituted in 50:50 TFE/100 mM ammonium bicarbonate (AmBic) (JTBaker) buffer; and, three frozen and three fresh samples were reconstituted in 0.1% Rapigest in 50 mM AmBic. Each sample was reduced in 10 mM of dithiothreitol (DTT) (Sigma-Aldrich) at 65ºC for 1 hr. To alkylate cysteine, samples were incubated for 1 hour at room temperature in 20 mM of iodoacetamide (IAA) (Sigma-Aldrich) in the dark. Enzyme digestion was carried out in 1:50 endoproteinase Lys-C (Promega) at 37ºC overnight followed by 1:25 Trypsin (Roche) digest at 37ºC for 2.5 hrs.

Samples treated with Rapigest reagent was quenched with formic acid (Sigma-Aldrich) to final concentration of 2% FA and incubated for 45 minutes at 37ºC. Samples were centrifuged at 16,000 g for 10 minutes, and then 50 µl (15 µg protein) were transferred to a clean tube.

All twelve samples were dried by a SpeedVac and stored at -80ºC until analysis.

Targeted Approach Preparation: For our targeted proteomic quantitation, samples were prepared as above, except in the end they were reconstituted in 0.1% formic acid in water (Honeywell B&J) and 20 fmol/µl of heavy (isotopically labeled) peptides of interest.

Blood-Spike Experiment: The optimal CSF sample preparation procedures were determined to be acetone precipitation followed by digest in Rapigest denatured samples

that did not undergo a freeze thaw step. This method was utilized in CSF samples spiked in with blood. Before acetone precipitation and digest, blood was spiked into pooled CSF at varying concentrations to examine the effects of blood contamination on CSF protein quantitation. The percent of blood added to each sample, and estimated hemoglobin concentration (µg/mL) is shown in **Table 1** below.

Sample ID	Blood added (‰)	Estimated hemoglobin conc. (µg/ml)
A	5.00	140.0
B	1.00	28.0
C	0.52	14.0
D	0.34	9.3
E	0.26	7.0
F	0.21	5.6
G	0.17	4.7
H	0.15	4.0
I	0.13	3.5
J	0.11	3.1

Table 1: Blood-Spike Experimental Protocol. Whole blood was diluted 1:1000 and spiked into CSF at varying concentrations (blood amount is shown in permille ‰). An estimation of hemoglobin concentration in µg/ml is listed for each sample.

Plasma: In order to enrich and be able to quantify low abundance proteins in plasma, additional steps were required, including depletion and addition of an internal standard (Enolase). Enolase (from Baker's yeast, Sigma) was added to ten plasma samples at varying concentrations from 0.1-90 µg/mL. To deplete the top 12 abundant proteins, samples were applied to Pierce™ spin columns (Thermo Scientific), capped, and mixed by Rotoflex (Argos) for 1 hr. at room temperature. Samples were placed in a collection tube and centrifuged at 1,000 g for 2 minutes and the column containing resin was discarded. For buffer exchange, about 500 µL of samples were added to an Amicon Ultra-

0.5 device (3000 kDa MWCO) and centrifuged at 14,000 g for 30 minutes (this step was repeated twice). To recover the solute, the Amicon was placed upside down in a clean micro-centrifuge tube and spun for 2 minutes at 1,000 g and the (~62 µL) recovered samples were reconstituted in 50 mM Ambic to 500 µL. Samples were then digested in 50% TFE. Each sample was reduced in 20 mM of dithiothreitol (DTT) (Sigma-Aldrich) at 65ºC for 70 min. To alkylate cysteine, samples were incubated in 20 mM IAA (Sigma Aldrich) at room temperature (in the dark) for 30 min. Enzyme digestion was carried out by adding endoproteinase Lys-C (Promega) 1:50 to total protein in each sample and incubating at 37ºC overnight. The next day, Trypsin (Roche) was added 1:25 of total protein, and samples were incubated at 37ºC for 2.5 hrs.

An additional experiment was completed for depleted plasma samples in 0.1% Rapigest followed by reduction, alkylation and dual digest as above. All samples were dried down by SpeedVac and stored at -80ºC for future analysis, however these sampleswere unable to run these samples due to instrument problems and high demand for use.

Mass Spectrometry Analysis

The sample analysis was completed on LC-MS/MS which is comprised of Q Exactive HF (Thermo Scientific) mass spectrometer coupled with nanoAcquity (Waters) HP LC. The separation was run on a C18 column (100 µm x 10 cm, 1.8 µm) column. The amount of total protein injection was 0.45 µg for unbiased and 1 µg for targeted analysis. A schematic of the instrument can be seen below in **Figure 3.**

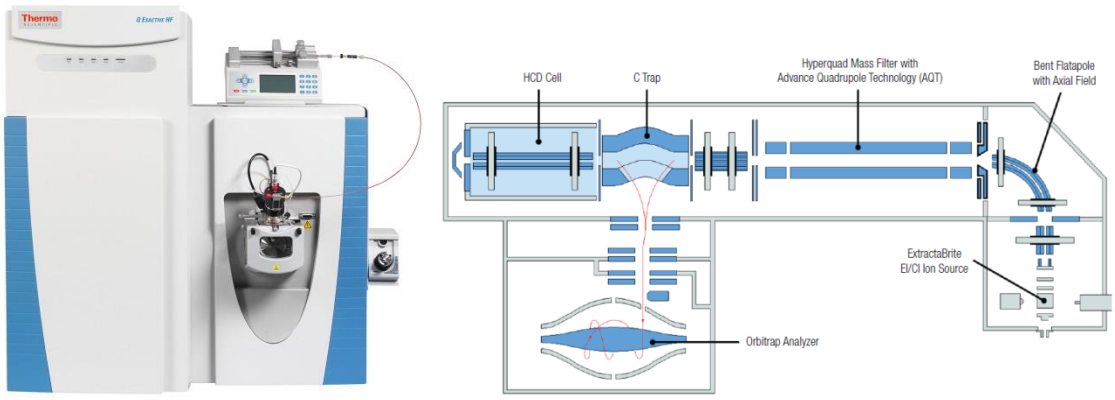


Figure 3: Schematic of Thermo Scientific™ Q Exactive™ HF Hybrid Quadrupole-Orbitrap Mass Spectrometer. The instrument utilized in this study for total protein quantitation after digest (and targeted spike in of heavy peptides) (Thermo Fisher, 2017).

Data Processing

Unbiased: Peak detection, alignment and feature quantitation was performed using Progenesis QI followed by peptide and protein identification in Mascot and Scaffold, respectively, using human UniProt and cRAP databases. False discovery rate for protein matching in Scaffold was set to 3% or lower. Matched proteins were reimported to Progenesis for relative protein quantitation and normalized to all proteins.

Targeted: Data was analyzed using Skyline, an open-source software for targeted proteomics analysis. A Spectral library for peptide identification was built from previously acquired unbiased data using Mascot data files. Peaks for the identified peptides were then quantified from areas of MS2 spectra and normalized against spiked-in heavy peptides.

RESULTS

The overall goal of this project was to establish an analytical method that allows the detection of low abundance proteins in a small volume of body fluids. The approach should maximize the number of proteins available for detection, be reproducible, and not significantly become affected by contamination from other fluids.

CSF:

Based on previous unbiased experiments in our lab (data not shown), we knew that acetone precipitation of CSF samples yielded ~50% more protein identifications than trichloroacetic acid (TCA) precipitation and better reproducibility. After acetone precipitation was confirmed as optimal, the variables of interest included the effect of introducing a freeze-thaw step, the digest method, and flow rate on the MS instrument. Each of these variables was tested in triplicate (three QC samples for all possible variable combinations). The following figures are representative of reproducible QC methods for digest method and introduction of freeze-thaw step.

Based on the strength of correlations (R^2 values) between QC samples in each condition, it was evident that the most reproducible method involved samples that were digested fresh (i.e. not frozen after precipitation) as well as those digested with Lys-C and Trypsin in the Rapigest surfactant (R^2=0.9730) (**Figure-4**). In contrast, CSF samples that included a freeze-thaw step and were treated with TFE detergent during digestion displayed the lowest amount of reproducibility (R^2=0.8114) **(Figure 5).**

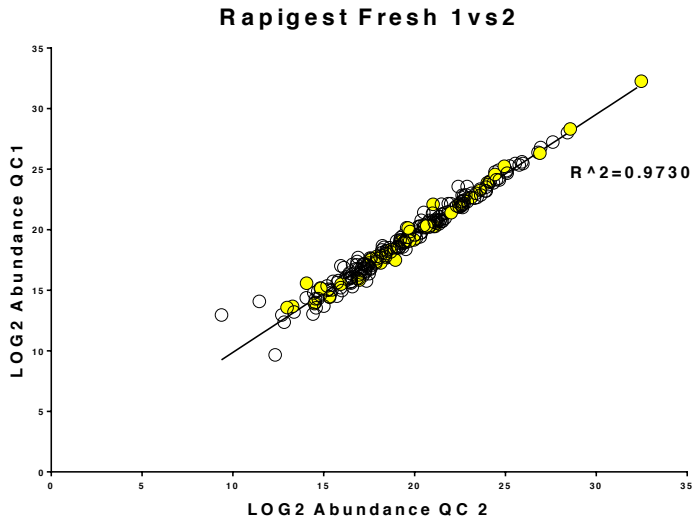


Figure 4: Representative Reproducibility Graph of Fresh Samples Using Digest in Rapigest Method. QC CSF samples were not introduced to a freeze-thaw step after acetone precipitation, and used Rapigest surfactant during digestion. This method was most reproducible (R^2=0.9730) compared to all other conditions and was used in subsequent CSF digest experiments. (Yellow circles represent peptides from targeted assay).

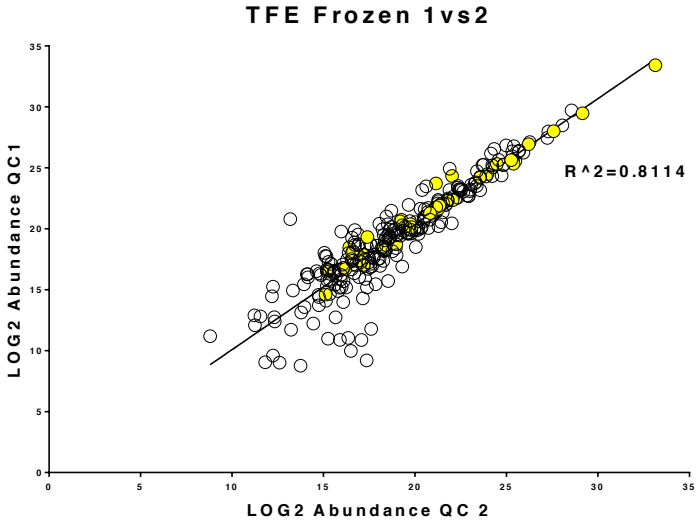


Figure 5: Representative Reproducibility of Frozen Samples Using Digest in TFE Method. QC CSF samples were introduced to a freeze-thaw step after acetone precipitation, and digested in 2,2,2-trifluoroethanol (TFE) reagent. This method was the least reproducible (R^2=0.8114) of all the other conditions tested.

Samples that were frozen and digested in Rapigest and those that were fresh and digested in TFE presented R^2 values of .9111 and .9314 respectively **(Appendix 1 and 2).** Therefore, in subsequent experiments regarding CSF peptide quantitation, we decided to use samples fresh, without a freeze thaw step, and digest with Lys-C and Trypsin in Rapigest reagent. Finally, each condition was analyzed by QEHF at high flow and nano flow. Nanoflow provided greater sensitivity and produced a greater number of protein identifications for all conditions (data not shown).

Blood Spike-In Experiment:

We were interested in spiking small amounts of blood into CSF to determine a) whether blood contamination affects peptide quantitations, and b) what percentage of blood causes the effect. Originally, we attempted to spike in a maximum of 5% of blood into small CSF volumes: our samples contained: 5%, 2.5%, 1.25%, 0.63%, 0.31% and 0.16% blood in CSF. The estimated hemoglobin (blood protein) concentration in these CSF samples would be 7 mg/mL, 3.5 mg/mL, 1.75 mg/mL, 0.88 mg/mL, 0.43 mg/mL and 0.22 mg/mL. Based on a blood contamination study by Hong et al., the cut off for protein quantitation effect is much lower at 200 ng/mL (Hong et al. 2010). We therefore adjusted our experiment accordingly to get a better understanding of whether blood contamination could affect our target list of peptide levels. We determined reasonable percentages of blood to spike-in by estimating hemoglobin concentration based upon the findings of Hong et al. We tested a gradient of estimated hemoglobin concentration and percent blood in CSF comparable to this minimum detection value. The results in light to heavy ratios of four representative peptides (from our 57 targeted peptides) compared to

that of a hemoglobin marker peptide is shown in **Table 2** below. The light to heavy ratio is a representation of the sample peptides (light) versus an added isotope labeled (heavy) internal standard. The heavy peptides contain the same amino acid sequence as the endogenous peptides, but with an added isotope label making it distinguishable which peptide is the standard and which should be quantified.

Sample Description			**Light/Heavy Peptide Ratio**				
Sample	**Est. Hemoglobin (μg/mL)**	**Blood (‰)**	**Peptide 1**	**Peptide 2**	**Peptide 3**	**Peptide 4**	**Hemoglobin Peptide**
A	140.00	5.00	7.06E-01	3.97E+00	6.63E-01	6.76E-01	1.27E+02
B	28.00	1.00	1.08E+00	1.92E+00	9.31E-01	1.05E+00	3.50E+01
C	14.00	0.52	9.16E-01	8.01E-01	9.16E-01	8.95E-01	1.35E+01
D	9.33	0.34	7.18E-01	8.26E-01	7.19E-01	8.04E-01	7.83E+00
E	7.00	0.26	9.50E-01	7.66E-01	9.55E-01	1.07E+00	7.55E+00
F	5.60	0.21	7.71E-01	8.29E-01	7.19E-01	8.75E-01	4.83E+00
G	4.67	0.17	7.08E-01	5.96E-01	6.60E-01	6.55E-01	3.49E+00
H	4.00	0.15	8.79E-01	5.99E-01	8.05E-01	8.80E-01	3.35E+00
I	3.50	0.13	7.61E-01	7.92E-01	7.32E-01	7.87E-01	2.98E+00
J	3.11	0.11	6.45E-01	4.10E-01	5.86E-01	6.08E-01	2.17E+00
K	0.00	0.00	1	1	1	1	1

Table 2: Blood Spike-in Study Results. Shown are the percent blood vs. light to heavy ratios for 4 chosen target peptides). Note that the amount blood is shown in permille (‰). Light to heavy ratios of five peptides of original 57 targets, including hemoglobin (blood) marker peptide for comparison. These four peptides are considered representative of remaining peptides quantified by MS/MS.

The data from the blood-spike in study is also visually displayed as a marked scatter plot **Figure 6.** A gradient of percentage (in per mille ‰) of blood spiked into each CSF sample is represented by the x-axis, and light to heavy ratio for each target peptide is on the y-axis. It is obvious that the blood spike-in shows strongly increasing levels of hemoglobin marker peptide (as expected), and based on this graph, only one peptide (peptide 2) was seemingly increased slightly at higher blood concentrations. Because

hemoglobin is a highly abundant protein in blood, the heavy to light ratio is expected to strongly increase with the amount of blood contamination as can be seen in Figure 6. However, since the increase in hemoglobin levels is so high, the Hemoglobin results in Figure 6 obscure any potential changes in the lower abundance normal CSF peptides (represented here by peptides 1-4).

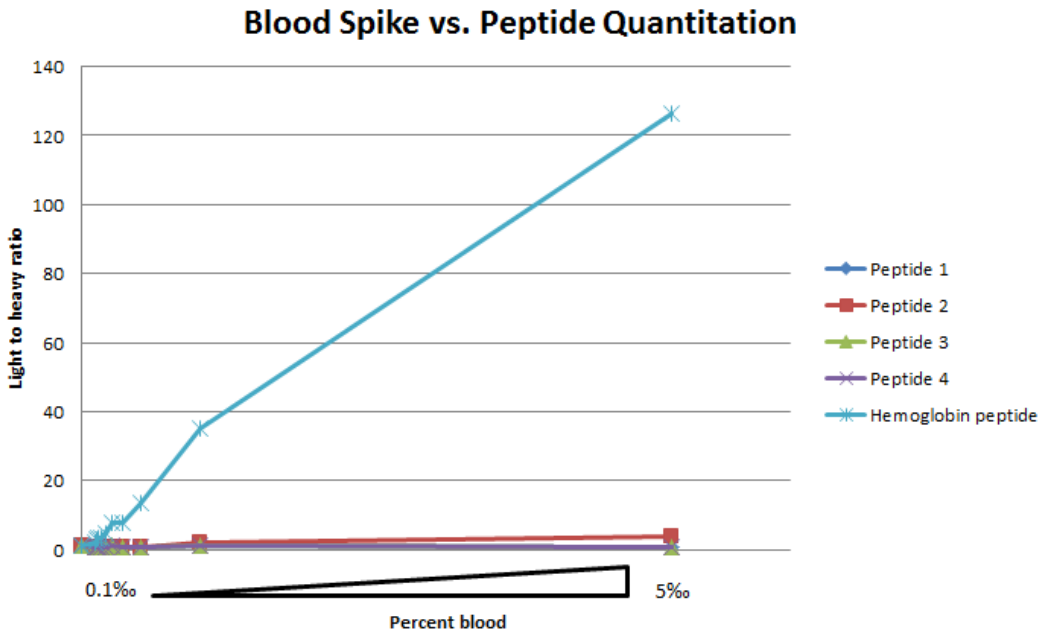


Figure 6: Effect of Increasing Blood Spike into CSF on 4 Target Peptides and Hemoglobin. Five peptides of interest from list of 57 targets are represented based on their light/heavy ratio due to blood contamination at various levels. Only hemoglobin peptide increased as blood contamination increased.

The data of the previous figure was re-plotted for clarity in **Figure-7** by expanding the Y-axis on a log base 10 scale**.** Based on this scatter plot, Peptide 2 levels appear to increase as the percent of blood contamination increases. Peptides 1, 3 and 4 are representative of the remaining 55 target peptides investigated in this project, and appeared to be unaffected by increasing amounts of blood contamination.

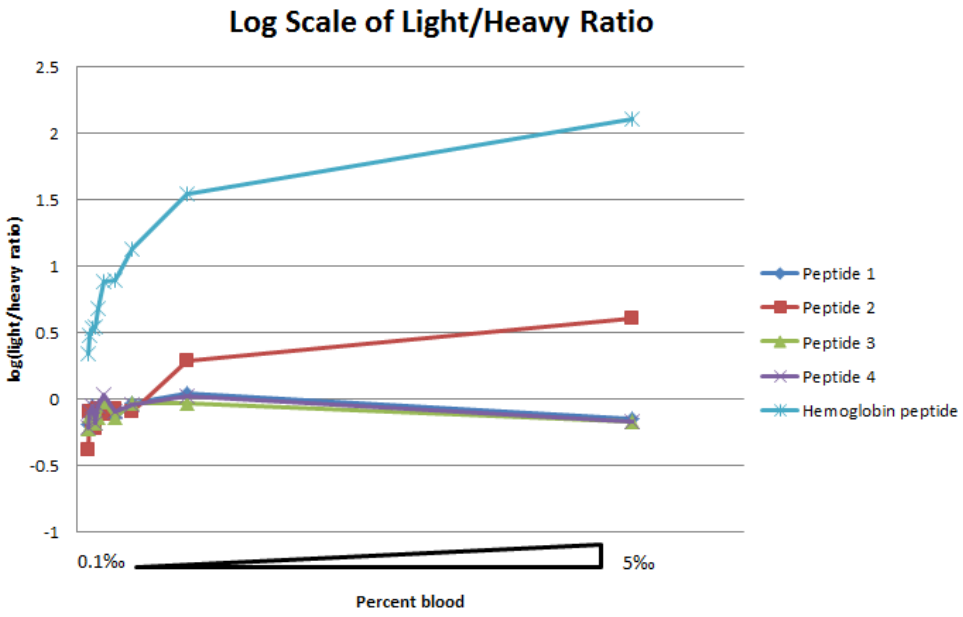


Figure 7: Marked Scatter Plot of Log10 Transformed Light/heavy Ratios from Blood Spike-in Experiment. Five peptides of interest from a list of 57 protein targets are shown based on the log of their light/heavy ratios at various blood concentrations. Peptide 2 (red) increased as the level of blood contamination increased.

Plasma:

Based on past optimization data, our lab determined that the best way to quantitate plasma peptides is to digest the samples in TFE reagent. We decided to utilize this technique in our experiments. As stated, plasma is a highly proteinated fluid containing a few high abundance proteins and many low abundance proteins. To prevent the high abundance proteins from interfering with the quantitation of the low abundance proteins, our group followed a previous protocol of depleting the high abundance plasma proteins before digesting. After digestion, an additional fractionation step was added before MS analysis to break up proteins even further into more easily identifiable

fractions and to reduce the complexity of the proteins. An internal standard (Enolase from Baker's yeast) was introduced into the plasma samples at varying concentrations from 0.1-90 μg/mL. Two of these experiments were completed with 9 samples each. However, due to time constraints, and instrument backups, the plasma samples were not analyzed by mass spectrometry for this report. Once completed, this experiment will provide information on the optimal amount of Enolase internal standard to be added during sample processing.

DISCUSSION

In determining the most optimal analytical methods for detecting low abundance proteins in small volumes of complex body fluids, the data in this project show that the optimum procedure depends on the matrix (body fluid) being tested. Unbiased MS methods are not highly reproducible, whereas using a targeted approach gives the advantage of focusing on quantifying specific proteins with good reproducibility. For unbiased quantitation in CSF, the best method of several tested in this project included use of acetone precipitation, no introduction of a freeze-thaw step, and digestion with Trypsin/Lys-C in Rapigest surfactant reagent. Although Trypsin/Lys-C digestion in TFE solvent yielded more unique protein identifications, most proteins were identified with only one unique peptide. Rapigest was chosen because it generated a higher percentage of protein identifications with 2 or more unique peptides, and the data was more reproducible than the TFE digest.

Collection of CSF by lumbar puncture is invasive, and bleeding can cause CSF sample contamination with blood. This contamination can cause misleading quantification data for specific target proteins when using CSF as a matrix to identify potential disease biomarkers. In our targeted study, we spiked blood at increasing concentrations into CSF samples to test for potential contamination effects. Our data indicated that blood contamination markedly affected hemoglobin peptide levels as expected, and that only one of our 56 chosen targeted peptides showed an apparent increase with increasing blood contamination. The effect was observed starting as low as 1.3‰ blood contamination in CSF. Therefore, blood contamination can artificially inflate observed MS peptide measurements and should be always taken into consideration.

A potential improvement of our methods might be to start with a larger volume of body fluid sample followed by the use of additional sample processing steps, such as enrichment, depletion and fractionation. However, due to the invasive nature of the collection procedure, CSF is not as readily available as other bodily fluids such as blood or urine, so this idea was not tested in this study. We could also have introduced an additional depletion step before CSF digest to remove higher abundance proteins as we did in the plasma analysis. The method that performed best in this study will be further optimized in additional future experiments.

BIBLIOGRAPHY

American Red Cross (2017) Learn about blood. Retrieved from: http://www.redcrossblood.org/learn-about-blood/blood-components/plasma on March 28th, 2017.

Anderson N, Anderson NG (2002) The human plasma proteome. *Mol. Cell. Proteomics* 2002, 1: 845–867. doi:10.1074/mcp.R200007-MCP200.

Chandrasekhar K, Dileep A, Lebonah DE, Kumari JP (2014) A Short Review on Proteomics and its Applications. *International Letters of Natural Sciences*, 17: 77-84. doi:10.18052/www.scipress.com/ILNS.17.77

Dayon L, & Kussmann M (2013) Proteomics of human plasma: A critical comparison of analytical workflows in terms of effort, throughput, and outcome. *EuPA Open Proteomics*, 1: 8-16. doi: 10.1016/j.euprot.2013.08.001.

Doerr A (2013) Mass spectrometry-based targeted proteomics. *Nature Methods,* 10(1): 23-23. doi: 10.1038/nmeth.2286.

Echan L A, Tang H-Y, Ali-Khan N, Lee K, & Speicher DW (2005) Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. *Proteomics*, 5(13): 3292-3303. doi: 10.1002/pmic.200401228.

Filip S, Vougas K, Zoidakis J, Latosinska A, Mullen W, Spasovski G, Mischak H, Vlahou A, Jankowski J (2015) Comparison of Depletion Strategies for the Enrichment of Low-Abundance Proteins in Urine. *PLOS ONE* 10(7): e0133773. doi:10.1371/journal.pone.0133773.

Gygi SP, Rist B, Griffin TJ, Eng J and Aebersold R (2002) Proteome analysis of low abundance proteins using multidimensional chromatography and isotope coded affinity tags. *Journal of Proteome Research*, 1: 47–54.

Hong Z, Shi M, Chung KA, Quinn JF, Peskind ER, Galasko D, Jankovic J, Zabetian CP, Leverenz JB, Baird G, Montine TJ, Hancock AM, Hwang H, Pan C, Bradner J, Kang UJ, Jensen PH, Zhang J (2010) DJ-1 and alpha-synuclein in human cerebrospinal fluid as biomarkers of Parkinson's disease. *Brain,* 133(3): 713-726. doi: 0.1093/brain/awq008.

Jacobs JM, Adkins JN, Qian W-J, Liu T, Shen Y, Camp DG II, Smith RD (2005) Utilizing Human Blood Plasma for Proteomic Biomarker Discovery *Journal of Proteome Research*, 4(4): 1073-1085. doi: 10.1021/pr0500657

Kellner R (2000) Proteomics. Concepts and Perspectives. *Fresenius' Journal of Analytical Chemistry*, 366(6): 517-524. doi: 10.1007/s002160051547.

Keshishian H, Addona T, Burgess M, Kuhn E, & Carr SA (2007) Quantitative, Multiplexed Assays for Low Abundance Proteins in Plasma by Targeted Mass Spectrometry and Stable Isotope Dilution. *Molecular & Cellular Proteomics*, 6(12): 2212–2229. doi:10.1074/mcp.M700354-MCP200.

Millioni R, Tolin S, Puricelli L, Sbrignadello S, Fadini GP, Tessari P, Arrigoni G (2011) High Abundance Proteins Depletion vs Low Abundance Proteins Enrichment: Comparison of Methods to Reduce the Plasma Proteome Complexity. *PLoS One*, 6(5): e19603. doi:10.1371/journal.pone.0019603.

Sickmann A, Dormeyer W, Wortelkamp S, Woitalla D, Kuhn W, & Meyer HE (2002) Towards a high resolution separation of human cerebrospinal fluid. *Journal of Chromatography B*, 771(1–2): 167-196. doi: 10.1016/S1570-0232(01)00626-2.

Tapiola T, Alafuzoff I, Herukka S, Parkkinen L, Hartikainen P, Soininen H, Pirttilä T. Cerebrospinal Fluid β-Amyloid 42 and Tau Proteins as Biomarkers of Alzheimer-Type Pathologic Changes in the Brain. *Archives of Neurology*, 2009; 66(3): 382-389. doi:10.1001/archneurol.2008.596

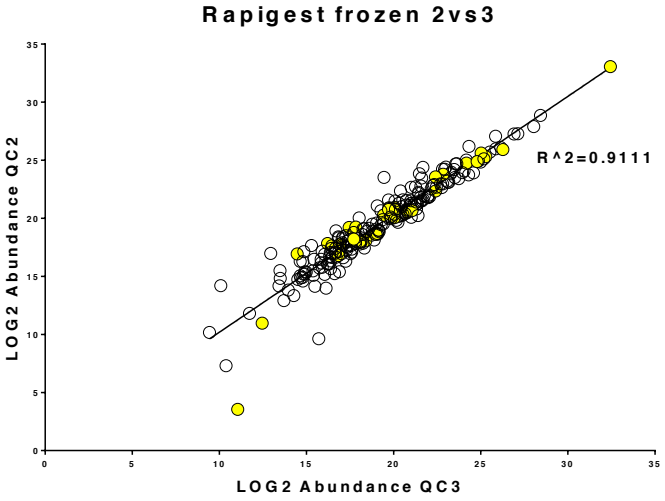
Thermo Fisher Scientific (2017) Overview of Mass Spectrometry for Protein Analysis. Retrieved from https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-mass-spectrometry.html.html on March 28, 2017.

Thermo Fisher Scientific (2017) Quantitative Proteomics. Retrieved from https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/quantitative-proteomics.html on March 28, 2017.

Zhang Y, Guo Z, Zou L, Yang Y, Zhang L, Ji N, Shao C, Sun W, Wang Y (2015) A comprehensive map and functional annotation of the normal human cerebrospinal fluid proteome. *Journal of Proteomics*, 119: 90-99. doi: 10.1016/j.jprot.2015.01.017.

APPENDICES

Appendix 1: Reproducibility of Frozen Samples Using the Rapigest Digest Method.



Appendix 2: Reproducibility of Fresh Samples Using TFE Digest Method.