Overview of a Common Data Analysis Pipeline for the Clinical Proteomic Tumor Analysis Consortium (CPTAC)

S.P. Markey, P. A. Rudnick, Y. I. Mirokhin, J. Roth and S. E. Stein*

National Institute of Standards and Technology

Mass Spectrometry Data Center

Biomolecular Measurement Division

Material Measurement Laboratory

* Contact: steve.stein@nist.gov

September 17, 2014

Index

Introduction

The National Cancer Institute (NCI) formed the Clinical Proteomic Technologies for Cancer Initiative in 2006 to address the pre-analytical and analytical variability issues that were major barriers to the field of proteomics. Based on the outcomes from the initial five years of research funding, an additional program was launched in August of 2011 entitled "The Clinical Proteomic Tumor Analysis Consortium", or CPTAC. CPTAC is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of robust, quantitative, proteomic technologies and workflows.

The participating proteome research groups in CPTAC recognized the need for a Common Data Analysis Pipeline (CDAP) in order to remove the multiple sources of variability that would result when trying to compare peptides and proteins inferred by each group using different software. This overview, for non-proteomic researchers, explains why and how processing choices produce results that appear to differ both qualitatively and quantitatively. If you already understand the elements of peptide and proteomic analyses using mass spectrometry, you might skip this document and proceed to use "*A Description of the CPTAC Common Data Analysis Pipeline (CDAP).*" That document details stepwise the software programs and output files of the Common Data Analysis Pipeline run at NIST.

Proteomics in a Nutshell

{ Note to the reader: If you are completely unfamiliar with proteomics, we suggest reading "What Is Proteomics". There are also numerous reviews which are indexed at PubMed.

If you are somewhat familiar with proteomics or are coming at this document from a related background (e.g., genomics), the following is a brief description and analogy as a refresher.}

Tissue samples are digested enzymatically to break large proteins into small segments (peptides containing 7-30 amino acids) that are amenable to automated analysis and assignment

of their amino acid sequences. The digests contain tens of thousands of peptides. Each tryptically digested tumor sample mixture is separated using multiple stages of chromatography to allow more effective mass spectrometric analyses of less complex mixtures. Typically, a patient tumor sample digest was chromatographically separated into 24 fractions, each of which was analyzed using high performance liquid chromatography (LC) coupled to a high-resolution tandem mass spectrometer (MS/MS). Once introduced into the mass spectrometer, a small mass region (containing 1 or a few peptides) is fragmented to produce a sequencing mass ladder where each peak in the spectrum (graph displaying m/z vs. relative intensity data) corresponds to a sub-peptide, fragmented with missing residues from one end or the other. It is these mass ladders that are analyzed by comparison to theoretical mass ladders produced by *in silico* digestion of a FASTA database by a search engine.

An analogy to the chromatographic separation and mass spectrometric characterization process is offered to convey the power of the analytical method and the complexities associated with reporting the results.

> *Suppose you anticipate a crowd of more than 100,000 people at a sporting event, and want to characterize that group to profile it relative to other crowds. To learn how many families are present, which children belong to which parents, how many attendees are male or female, how many have red/brown/black/grey/no/other color hair, which ones are related by marriage, which ones share religious beliefs, as well as other details of their personal lives, it would be helpful to pass them through a turnstile and profile each of them individually as they enter the stadium.*
>
> *For protein digests, the chromatographic step acts like an imperfect turnstile, occasionally separating peptides as individuals, but more frequently allowing small groups*

of 5-50 peptides to enter together. The tandem mass spectrometer analyzer acts like a ticket taker and a discriminating usher, dividing peptides first by size (actually, mass-to-charge or 'm/z' ratio) and then by a host of their secondary characteristics (actually, 'MS2 spectrum'). The details that emerge from the MS2 spectrum allow each individual peptide to be distinguished and characterized as being related or unrelated to others in the mixture.

Why a Common Data Analysis Pipeline?

A continuation of this analogy is given in support of a common data analysis pipeline.

So how can there be different lists summarizing the characteristics of all of the individuals present in the stadium or proteins in a mixture? Well, suppose that the stadium data was analyzed by crackerjack polling teams and statisticians from both The Wall Street Journal and The Washington Post. The same raw data would be coded and entered for identical individuals, but two very different profiles of the stadium crowd could emerge. Both summaries could be valid, and not easily reconciled by reading the resulting reports. One statistical team might use home address plus cell phone telephone numbers plus 2012 voter registration lists to designate their familial groupings; the other might use zip code plus landline telephone numbers plus 2010 motor vehicle registration lists from several states for the same purpose. The analogous software tools that proteomics specialists choose have many of the same characteristics that make reconciliation of a stadium crowd's characteristics difficult when <u>different dated, time-dependent, and incomplete reference databases</u> are used. Why is this true? Some stadium attendees (babies-toddlers) were born after the databases were compiled; some people moved or married after the lists

were finalized; some have changed their phone numbers; some are visiting from Europe to attend a family reunion; others are visiting from Australia and are unrelated to anyone else in the stadium. Analogous problems arise in every proteomics investigation. It is impossible for statistical professional polling teams to accurately align individuals and family relationships after the fact, whereas it might have been obvious to an attentive and very inquisitive usher at the time of seating. Correct assignment of peptides to their parent proteins is fraught with all of the problems associated with placement of individuals into correct family relationships. While some are unambiguous (distinct peptides that can only belong to a single protein record, coded by a single gene), others are shared, common to multiple related proteins coded by several genes. The resulting quantitative assessment of a stadium crowd or a protein list becomes dependent upon the evaluator and the evaluator's measurement tools. This becomes even more apparent when time has passed and the same evaluators profile another stadium crowd or peptide list, but the databases for their analyses have changed, so that the resulting lists have many non-comparable entries. In contrast, it is certainly possible to compare a stadium crowd at a football stadium with an equivalent sized crowd at a baseball stadium many months later if the same evaluators apply the same sets of tools to the crowd characterization. For this reason, the NCI determined that a common data analysis pipeline would be beneficial for reporting results from multi-institutional cancer tumor proteome studies.

Is Original Instrument Data Retrievable from the Data Coordinating Center?

Data resulting from tandem mass spectrometry analyses is recorded in electronic digital files and stored in formats that are instrument-specific, unaltered over time, and easily and accurately replicated. The Proteome Characterization Centers (PCCs) at Vanderbilt University, Broad Institute, Pacific Northwest National Laboratory (PNNL), and Johns Hopkins University(JHU) used Thermo Fisher mass spectrometers, albeit of several different models, but producing similarly formatted primary data, denoted as 'Filename.raw' files. These Filename.raw files were transferred from each participating lab to a central Data Coordinating Center (see Figure 1).

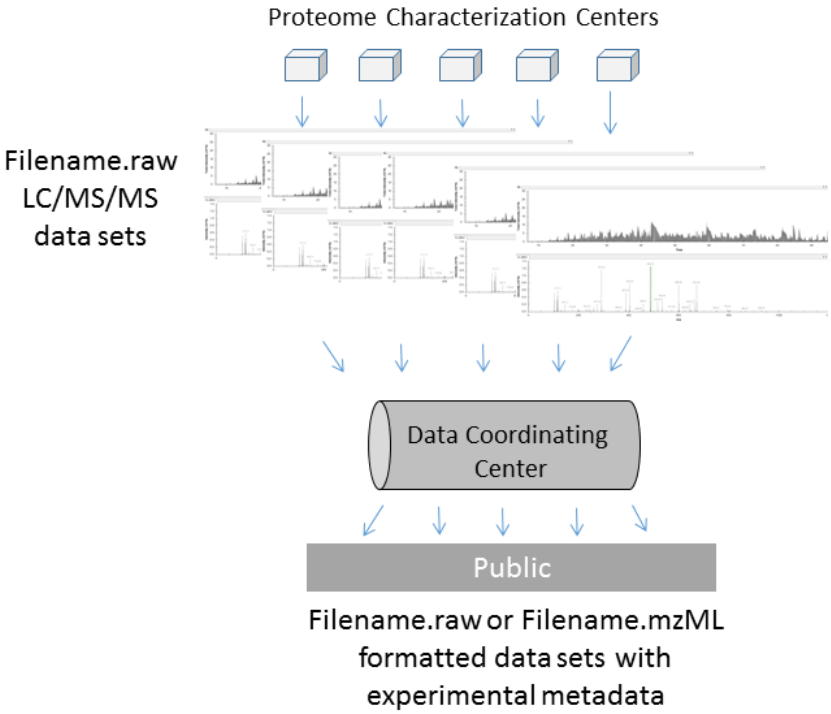

Figure 1. Publicly accessible native and readable LC/MS/MS experimental data

All Filenames were chosen to follow a standardized 7-part naming convention, described in a

document entitled "CPTAC, TCGA Cancer Proteome Study of Breast Tissue Naming Conventions" available on the Data Coordinating Center website. Primary Filename.raw files are available for public download that preserve the original quality of all the recorded experimental data. However, to facilitate review of this data for those without access to Thermo Fisher proprietary software, the Filename.raw files were converted to 'Filename.mzML' files. The Filename.mzML formatted files can be viewed with open source ProteoWizard software tools. Alternatively, the original instrument Filename.raw files can be converted to mzML or other ASCII formats using those same tools, following the installation of MSFileReader from Thermo Fisher.

How Are Lists of Peptides and Their Intensities Generated at NIST?

The processing of the original Filename.raw instrument files begins with conversion of the data from 'profile mode' (point-by-point detail that includes m/z peak shapes, resolution, and noise) to 'centroid mode', simple peak lists of m/z vs. intensities (Filename.mgf).

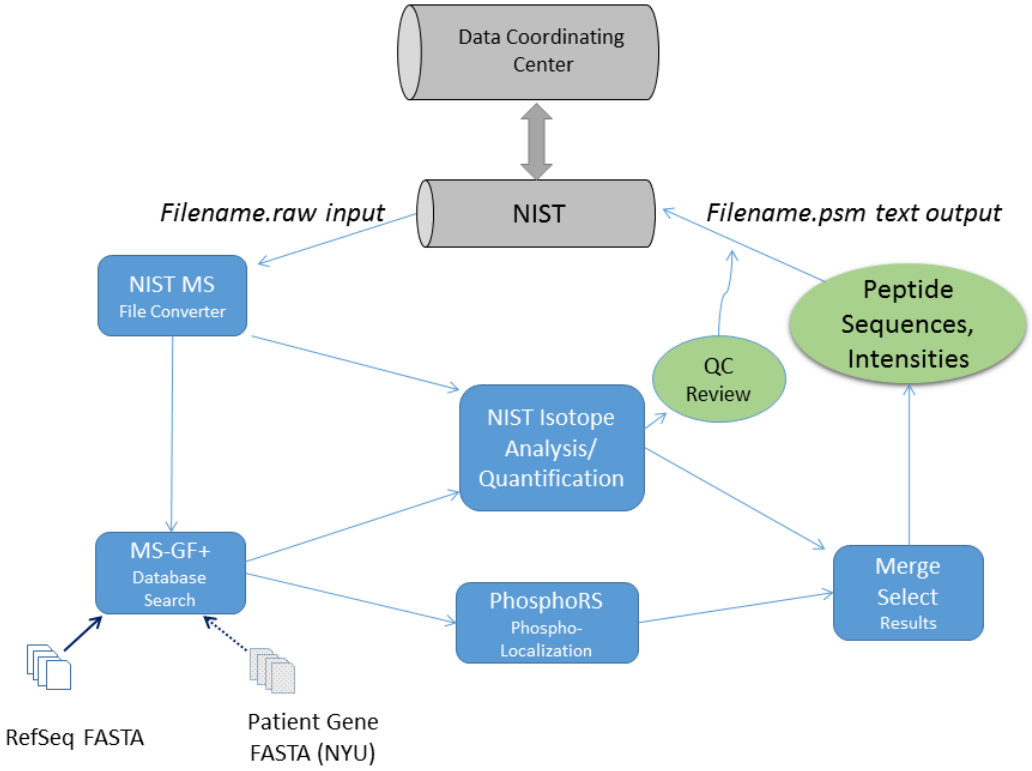

Figure 2. NIST pipeline for processing LC-MS/MS data to peptide sequences and intensities

The raw data is preserved in text format profile mode for isotope analysis and quantification software (Filename.mzXML). The file conversion processes use a NIST expanded version of the ReAdW converter software from ISB ReAdw4Mascot2. The peak list file (Filename.mgf) is annotated to include parameters important to peak area quantification in iTRAQ experiments used by Broad, PNNL, and JHU. The iTRAQ experiment requires each

sample to be reacted with a chemical label, and as a result, allows 3 patient samples to be mixed and analyzed together with a pooled reference standard employed for all samples. The iTRAQ experimental results contain sets of quantitative ratios for each set of 3 patients with respect to the common pooled standard. The iTRAQ process introduces multiple subtle factors into data analysis, and consequently, the Filename.mgf record includes notations for peak purity and missing peaks. [For a concise description of the iTRAQ quantification method used in these analyses, view this link.]

The next steps are to assign peptide sequences to each MS2 spectrum and perform quantitative analyses (Figure 2). There are many software tools for peptide sequence assignment, but NCI and NIST selected MS-GF+ after testing several alternatives. MS-GF+ requires a database of protein sequences in a standardized text format (FASTA) in which each amino acid is represented by a single letter code. The database used in processing the TCGA samples is the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) for Homo sapiens (build 37) including the sequence for S. scrofa (porcine) trypsin added to each sample. The RefSeq database is a non-redundant collection of protein sequences from archival databases. A second database compiled from TCGA data for individual patient genes is being evaluated for possible re-analysis with MS-GF+. NIST developed software for peak profile isotope analysis and quantification (ProMS) that uses both the Filename.mzXML and Filename.mzid input to produce intermediary and non-public working files (Filename.txt).

What Types of Analyses Were Performed on Each Tumor Type? Are They Directly Comparable?

Tryptically-digested samples can be divided for the purpose of two different types of analysis – as peptides, and/or as post-translationally modified peptides. The PNNL and Broad groups chose to separate and characterize both peptides and post-translationally phosphorylated (addition of a phosphate on serine, threonine or tyrosine hydroxyl groups) peptides as potentially characteristic of ovarian and breast tumor biology. JHU chose to analyze peptides and separate glycosylated peptides for ovarian cancer samples; Vanderbilt did not split the colon cancer samples for post-translational peptide analysis, but added a large sample set of normal colon tissues.

The interpretation of mass spectra of phosphorylated peptides requires analysis in addition to MS-GF+ in order to assign the likely position(s) of phosphate group attachment, and an assessment of the probability of that assignment. The software program PhosphoRS [Taus *et al*., *J Proteome Res.* 2011;10(12):5354-62] provides that information, and was incorporated into the pipeline (Figure 2).

In contrast, JHU trapped N-glycosylated peptides onto an ion exchange column, and enzymatically cleaved the asparagine-linked peptides prior to elution. Consequently, the resulting de-glycosylated peptides can be sequenced using MS-GF+, with the expectation that the former glycosylated asparagine residues will be identified as aspartic acid.

Although the same general technology (liquid chromatography-tandem mass spectrometry) was used by all of the labs, the results should be expected to differ because of the above differences in sample processing. The iTRAQ-labeled peptides will differ from those analyzed

by label-free global analysis. The iTRAQ/phosphoproteomics will differ from those after de-glycosylation/iTRAQ analysis. Each specific analytical method would be expected to reveal slightly different characteristics of the tumor samples, with each one having merit, and all likely to yield greater information than any one. One set of samples was analyzed in part by two laboratories using different techniques. Both PNNL and JHU analyzed 32 ovarian cancer tumors in common using an iTRAQ/global +phospho method (PNNL) and an iTRAQ global + deglycosylation method (JHU).

How Was Quality Control Measured? Were Standard Reference Materials Used?

NIST performs quality assessment using parameters derived from each of the output files from quantitation and isotope analysis[Rudnick *et al., Mol Cell Proteomics* 2010; 9(2):225-41]. The files are reviewed as complete sets of runs so that changes in sample handling, instrument performance, chromatography, or computer data handling will be detectable. Examples of measured metrics parameters are shown in Figures 3 and 4. NIST quality control programs calculate and track more than 40 system characteristics, and four of those that reflect chromatography and mass spectrometry instrument performance are shown in Figure 3. The number of identified tryptic peptide sequence matches (Figure 4) captures and reflects the bottom-line performance consistency for any given laboratory.

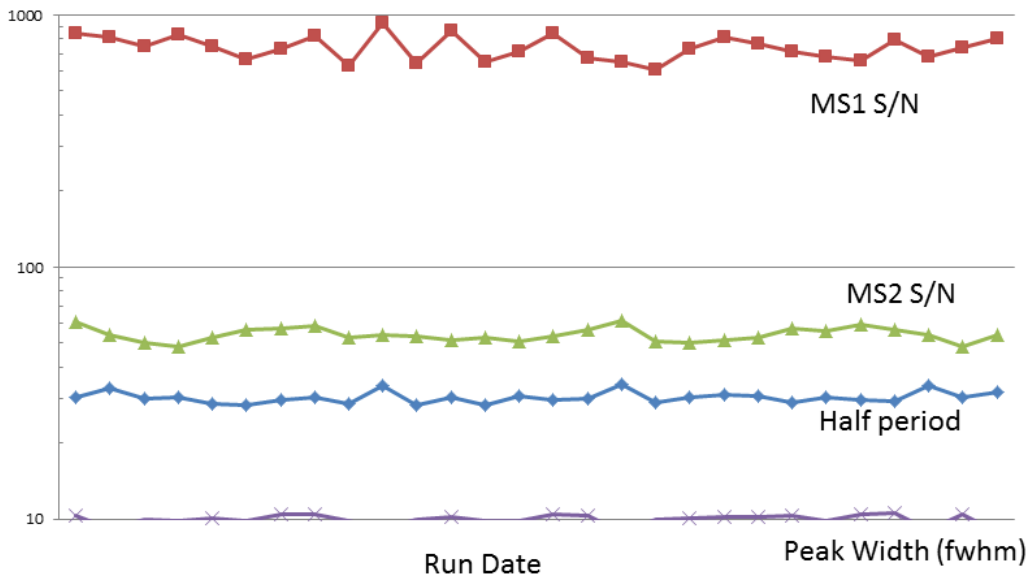

Figure 3. Example of NIST LC and MS Instrument QC Metrics(means)

Figure 3. illustrates the consistency of several different performance metrics for 28 separate analyses collected at one of the Proteome Characterization Centers. MS1 S/N is the signal-to-noise ratio measured for *m/z* data after the first stage of orbitrap high-resolution mass analysis; MS2 S/N is the same metric after the second stage tandem analysis. Half-period refers to the time over which the middle 50% of the identified peptides elute, and peak width (fwhm) is a measure of whether the chromatographic peaks are changing over time. Both metrics describe chromatographic separation quality and consistency.

The result of NIST QC metric monitoring was an observed consistency within each laboratory for analyses performed of the TCGA samples. Because each laboratory conducted internal performance checks, and re-ran samples when necessary, there were no failed patient sample runs found in the data pipeline. This consistent performance level was possible because

each participating laboratory pre-tested their experimental protocol with system suitability studies using human-in-mouse xenograft breast cancer tumor reference material (CompRef) distributed to all groups for lab-to-lab and within-laboratory performance checks. The same CompRef materials were run between TCGA samples for quality control, and the resulting 'interstitial' CompRef analyses are available for download on the DCC site.

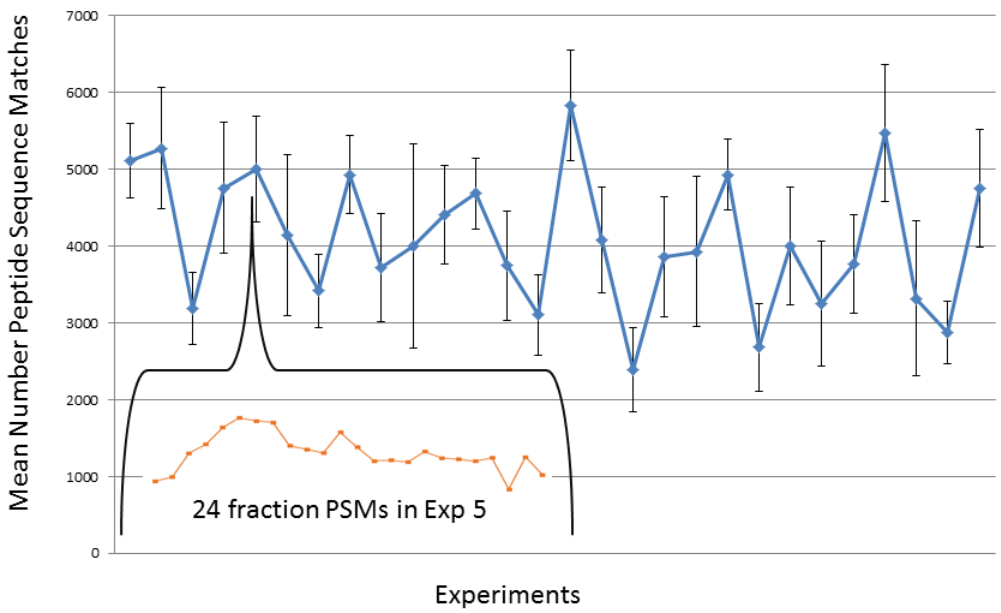

Figure 4. NIST Metric Example of Peptide Sequence Matches Across Experiments, Each Containing 24 Fractions

Figure 4 tracks the number of identified peptides across different experiments from one PCC. Each point is a mean of the number of peptides (±SD) identified from 24 fractions, as illustrated in the inset graph for Experiment 5. Typically, a lower number of peptides elute in the early and late fractions, and a spread of values is expected due to biological variations. Monitoring numbers of peptide sequence matches is a measure of overall PCC performance.

How Were Proteins and Genes Assigned?

Each peptide sequence in the Filename.psm text report is linked to the list of proteins that contain that sequence in the reference database, and as a consequence, an all-inclusive list of all possible proteins can be derived from those reports. However, compiling an all-inclusive list of all possible proteins is not very useful to biologists because it violates the principle of parsimony (Occam's razor). Applying this principle, the shortest list of candidate proteins or genes that can explain all of the data is more likely to be correct. Software for performing parsimonious protein assignments requires that peptide sequences and candidate proteins be considered and sorted together in order to solve for the smallest set solution (Figure 5).

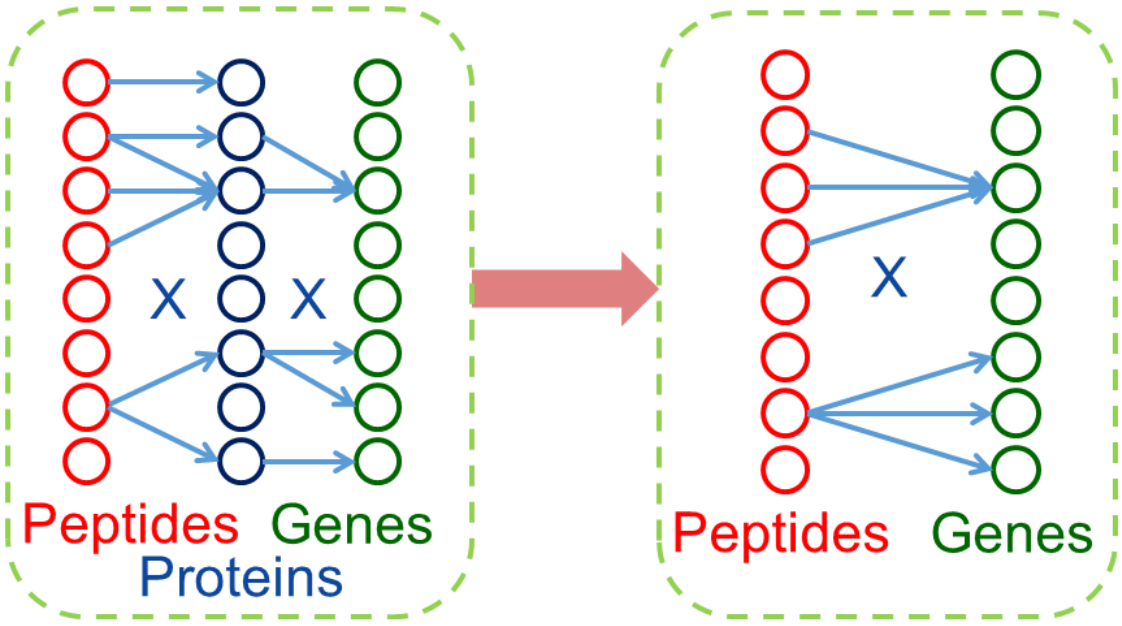

Figure 5. Gene based inference (2 peptides required) avoids complications arising from an intermediate protein inference step

N. Edwards

Compounding this task, there are many shared peptides among protein isoforms. Consequently, it is not possible to determine quantitatively how much of each peptide originated from a specific

protein isoform. Because biologists view proteins as gene products, we elected to bypass the protein isoform conundrum and assign peptides directly to a parsimonious set of genes (Figure 5), a task performed at DCC by Dr. Nathan Edwards using software designed for this purpose.

While gene assignment does not eliminate quantitative assignment ambiguity for all cases, the inferred parsimonious set of genes generates a simpler path to the desired output needed by biologists and medical professionals, particularly those assessing biological networks and systems. For biologists interested in biomarker candidates, the only meaningful experimentally measured quantitative data resides in the peptide sequence match (Filename.psm) reports. Any inferred summation of quantitative peptide data necessitates compromises.

What Type of Gene Summary Reports are Available?

The gene summaries are experiment specific, so that there are separate sets for PNNL-ovarian, JHU- ovarian, Broad breast, and Vanderbilt colo-rectal cancers. Each experiment has text summaries, gene-inference; peptide-gene relationships; gene-inference based quantification based on iTRAQ ratios using m/z117 as a pool reference value; and for label-free (Vanderbilt) gene-inference based spectral count and precursor area quantitation.

Why Are Mass Spectral Library Spectra Produced?

How Can They Be Accessed?

Mass spectral files accumulated by the CPTAC project contain >100 million mass spectra. The mass spectrum of each unique peptide sequence exhibits a characteristic reproducible pattern of mass/charge vs. intensity, much like an individual's fingerprint. Consequently, mass spectral libraries of previously characterized components permit very rapid

identification of the same peptides when encountered in future studies, a process not unlike finding facts in a library of bound volumes, as illustrated in Figure 7.

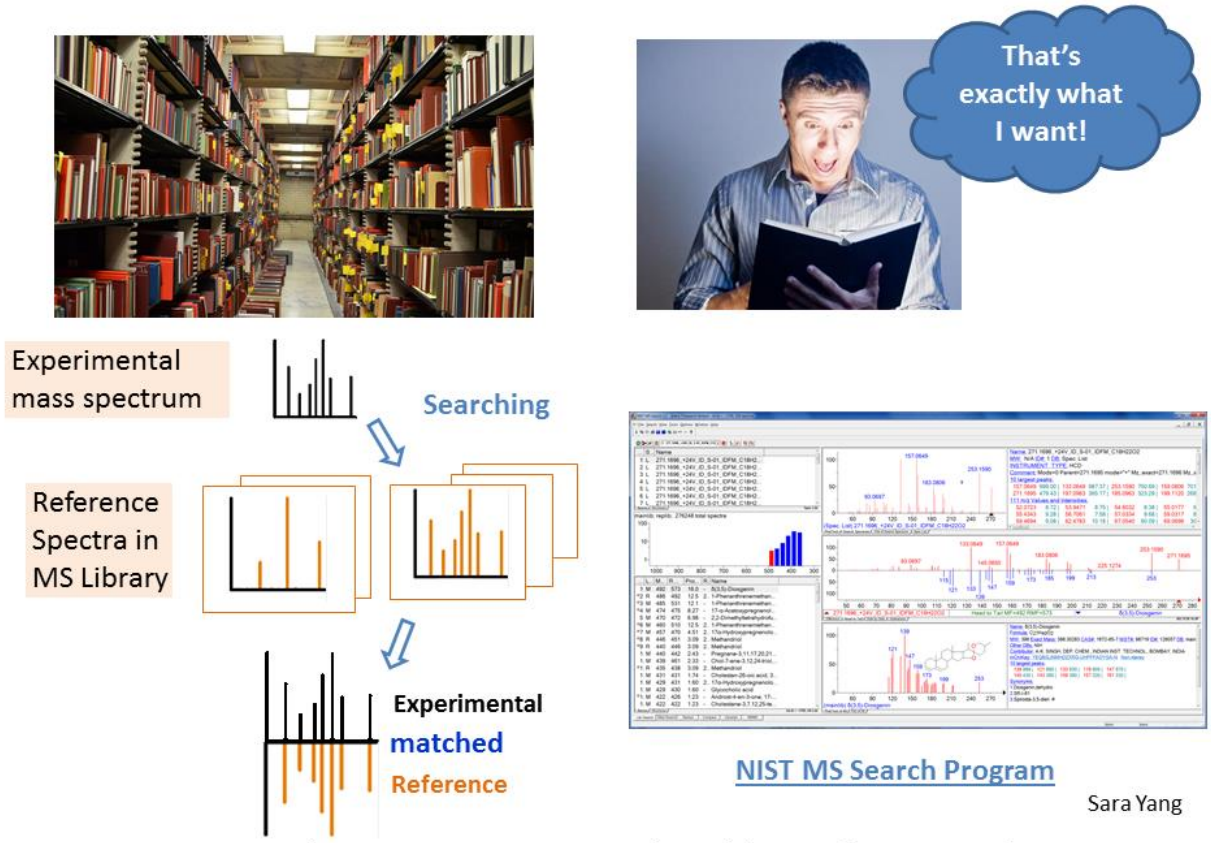

Figure 6. Mass spectral pattern recognition is similar to library reference searching, but markedly accelerated with software tools.

The NIST Mass Spectrometry Data Center established repositories of compound specific mass spectral data useful for rapid recognition of simple chemical structures like drugs, pesticides, steroids, amino acids, etc., beginning in the 1970s. These libraries and associated software enabling spectral matching have been widely accepted in analytical laboratories worldwide. More recently, libraries of tandem mass spectra of peptides recorded using liquid chromatographic separation, electrospray ionization using ion trap-type instrumentation have been distributed to the public by NIST after several steps of curation. Composite consensus

spectra are derived from comparing many spectra of the same peptide determined at different intensities, as illustrated in Figure 7.

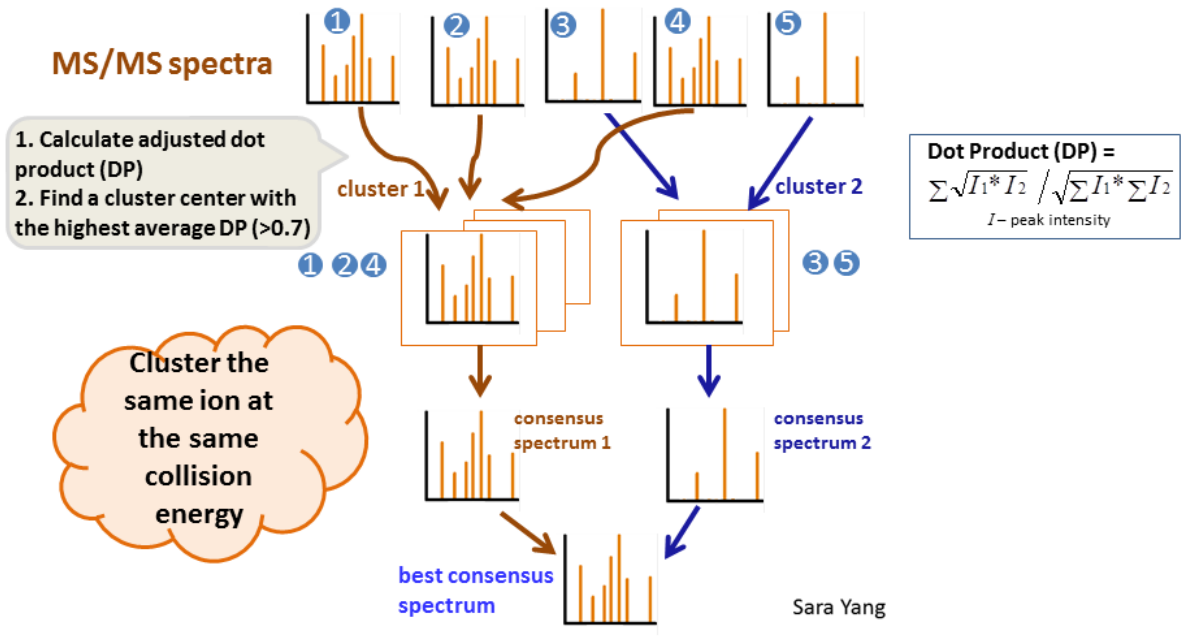

Figure 7. Consensus mass spectra are derived from the many spectra all assigned to the same peptide, but differing in intensities and chemical noise background. The process of clustering similar spectra leads to the generation of a composite spectrum, containing minimal chemical background and noise.

Compilations are assembled from spectra acquired using similar mass spectral instrumentation. For example, the Vanderbilt data yielded a library of > 98 thousand consensus spectra. These were merged in the latest public release of the NIST Human tandem peptide library that now contains >340 thousand consensus tandem spectra recorded using ion trap instrumentation. Data from the Broad, JHU, and PNNL studies was collected using a cycloidal ion trap with image current detection (Orbitrap ™), and resulted in a library of >1 million consensus iTRAQ spectra distilled from > 56 million total collected spectra. Similarly, separate libraries were compiled

from global and iTRAQ-Phospho Orbitrap™ data sets. All of these libraries are being distributed through the NIST public websites or through links accessible on the CPTAC-DCC website.

Overview of Processed Data

What Makes Files from each Institution Unique?

The CPTAC studies of TCGA samples were planned to utilize state-of-the-art instrumentation in each institution with the expectation that results are parallel, but not precisely mirrored.

Table 1

	TCGA CA Sample	Instrument/ Analysis	MS/MS Mode	Separation (fractions)	Chromatography
Broad-MIT	Breast	Thermo Q-Exactive/ iTRAQ (4-plex)	HCD	SCX (24 global + pool 12 phospho + pool)	C18, 75 µ 120 min
Johns Hopkins	Ovarian	Thermo LTQ Orbitrap Velos/ iTRAQ (4-plex)	HCD	Basic pH-RP (24 global + pool; single glyco fraction, 3x)	C18, 75 µ 90 min
PNNL	Ovarian	Thermo LTQ Orbitrap Velos/ iTRAQ (4-plex)	HCD	Basic pH- RP (24 global 12 phospho)	C18, 75 µ 100 min global 175 min phospho
Vanderbilt	Colo-Rectal	Thermo LTQ Orbitrap Velos	CID	Peptide IEF (15)	C18, 100 µ 95 min

Table 1 summarizes some of the key factors that are similar, but differ not only with respect to tumor type, but also the sample workup protocol, analytical instrumentation, and separation.

How differences in protocols are likely to affect results can be anticipated from general understanding of the experimental options. Most importantly, the protocol differences should not alter the underlying biological conclusions reached. The iTRAQ protocol should produce tighter quantitative analytical data for comparisons between samples. Think of competitive track racing for an analogy for 4-plex iTRAQ experiment. If 4 runners compete in a track race together and the winner of the first race competes with another set of 3 runners, the result will allow precise comparison of the entire set of seven in the time required to run two races, similar to an iTRAQ experiment where 3 samples are mixed with a 4^{th} as a consistent standard. In contrast, if seven athletes compete by running individual heats, the seven events will require longer, in analogy to the non-iTRAQ method used by Vanderbilt. The trade-off is that observers of both events view runners either in a group or as individuals. The iTRAQ experiment somewhat dilutes the signal for each component, but the non-iTRAQ alternative increase comparative quantitative variability and requires longer instrument time. Other differences in the protocols (numbers of fractions, type of columns or fragmentation) are technical, like differences in track surface or weather, and not likely to produce substantive differences in data sets.

What Makes All Data Sets Comparable?

The use of a Common Data Analysis Pipeline results in files that can be directly queried and compared with respect to peptides and genes that may be indicative of activated or suppressed pathways in different cancer tumor types. The parameters listed in Table 2 define some of the many options that were applied consistently to the processed data sets. Within proteomics, there are many data analysis software tools, and, in addition to the CDAP, multiple analyses of CPTAC data appear in the scientific literature. It is likely that summary data

published by each institution differs qualitatively and quantitatively somewhat from that processed by CDAP because each lab reports using specifically dated reference databases and software tools.

	CDAP	Broad	JHU	PNNL	Vanderbilt
FASTA	RefSeq-Human-v37-Trypsin.fasta (32,800 entries)	Same as CDAP	Same as CDAP	Same as CDAP	humanRefSeq_v54_trypsin.fasta (34,589 entries)
Search Engine(s)	MS-GF+ (v9733)	SpectrumMill 4.0 (Beta)	MS-GF+ (v9146)	1. MS-GF+ v9324 (2/27/2013) v9358 (3/05/2013) v9593 (05/06/2013) v9699 (07/26/2013) v9736 (09/16/2013)	1. Pepitome 1.0.42 (library) 2. MyriMatch 2.1.87 3. MS-GF+ (v9176)
Ambiguous matches flagged?	Yes	No	No	Yes	Yes
Variable Protein Mods searched	MetOx(+16) Deamidation(+1)	MetOx(+16) Glu->pyro-Glu(-18) Gln->pyro-Glu(-17) Deamidation (N)(+1)	MetOx(+16)	MetOx(+16)	MetOx(+16) Glu->pyro-Glu(-18) Gln->pyro-Glu(-17) Acetylation (+42)
Semi-tryptic searched	Yes	No	Yes	Yes	Yes
Precursor tolerance	20 ppm	20 ppm	10 ppm	10 ppm (post-DTARefinery)	20 ppm
Missed Cleavages	No limit	<5	<2 post search	No limit	By search engine
False Discovery Rate	1% PSM	1% PSM	1 % peptide	1% Peptide	1% PSM

Table 2

How and Why Institute Published Data May Differ from CDAP Results

First, the bioinformaticians at each Proteome Characterization Center select the search engines, reference databases, and parameters that they believe will produce the most useful and comprehensive data analysis for their output. While a committee of Proteome Characterization Center members agreed on the publicly accessible and well documented tools and methods for the common pipeline, the same scientists were free to select software and modify/extend databases where they thought there were significant advantages for processing their own data. For example, the team at PNNL first 'refined' their mass spectral data, having recognized that certain instrumental drift could be eliminated while improving the accuracy of m/z assignments. Subsequent peptide searching using the same MS-GF+ software enhanced the numbers of peptide sequence matches from data relative to CDAP results. In contrast, the JHU team selected conservative parameters for processing that yielded slightly fewer peptide sequence matches using the same MS-GF+ software. The Broad team prefers a different search engine (SpectrumMill, proprietary software from Agilent). Their results substantially agree with those from the CDAP, but the CDAP's allowance of semi-tryptic peptides resulted in some peptide sequence matches that the Broad team prefers not to include for iTRAQ quantification. Vanderbilt's bioinformatics team elected to use multiple search tools (library search with 'Pepitome' and a second search engine 'MyriMatch') to improve their ability to assign spectra relative to the single MS-GF+ search used by CDAP. The multiple search engine and library strategies are well documented to enhance peptide sequence matching for the global label-free proteomics using Collisional Induced Dissociation (CID) fragmentation in the LC/MS/MS experiment performed at Vanderbilt. Library search strategies are not applicable to the Higher-

energy C-trap Dissociation fragmentation (HCD) data produced at the other centers, but all of the other TCGA data could yield a greater number of peptide sequence matches with the use of additional search engines that employ strategies orthogonal to that of MS-GF+.

TCGA Proteome Data Sets are a Rich Resource for Bioinformatics Investigations

The composite data sets for each TCGA sample experiment are larger and more comprehensive than any previous proteomic investigations of cancer tumors. Consequently, they are useful for future bioinformatics analyses on many levels. Additional processing is planned both at NIST/DCC and at each Proteome Characterization Center. For example, patient-specific genome data will be used to search all data sets for tumor specific peptides. At the same time, employing second or third search engines may enhance peptide sequence detection. There are large numbers of unidentified MS2 spectra from abundant components (30% or more of the total acquired spectra) that are consistently observed in samples, and these require investigation and characterization. NIST libraries of frequently encountered, unidentified mass spectra will aide future research by rapidly recognition of novel components vs. commonly encountered artifacts.

As with any data analysis, the CPTAC data require careful consideration of the assumptions and uncertainties inherent within the analysis methods and measurements. The NIST team is responsible for ensuring that biological and clinical conclusions drawn from CDAP data are based on a consistently high level of proteomics data quality.